arXiv:2507.19593v1 [cs.AI] 25 Jul 2025

Hypergames: Modeling Misaligned Perceptions and Nested Beliefs for Multi-agent Systems

Vince Trencsenyi[1], Agnieszka Mensfelt[1], Kostas Stathis[1]

[1]Department of Computer Science, Royal Holloway University of London, Egham Hill, Egham, TW200EX, Surrey, United Kingdom.

Contributing authors: vince.trencsenyi@rhul.ac.uk; agnieszka.mensfelt@rhul.ac.uk; kostas.stathis@rhul.ac.uk;

Abstract

Classical game-theoretic models typically assume rational agents, complete information, and common knowledge of payoffs – assumptions that are often violated in real-world multi-agent systems (MAS) characterized by uncertainty, misaligned perceptions, and nested beliefs. To overcome these limitations, researchers have proposed extensions that incorporate models of cognitive constraints, subjective beliefs, and heterogeneous reasoning. Among these, *hypergame theory* extends the classical paradigm by explicitly modeling agents' subjective perceptions of the strategic scenario, known as *perceptual games*, in which agents may hold divergent beliefs about the structure, payoffs, or available actions. We present a systematic review of agent-compatible applications of hypergame theory, examining how its descriptive capabilities have been adapted to dynamic and interactive MAS contexts. We analyze 44 selected studies from cybersecurity, robotics, social simulation, communications, and general game-theoretic modeling. Building on a formal introduction to hypergame theory and its two major extensions – hierarchical hypergames and the hypergame normal form (HNF) – we develop agent-compatibility criteria and an agent-based classification framework to assess integration patterns and practical applicability. Our analysis reveals prevailing tendencies, including the prevalence of hierarchical and graph-based models in deceptive reasoning and the simplification of extensive theoretical frameworks in practical applications. We identify structural gaps, including the limited adoption of HNF-based models, the lack of formal hypergame languages, and unexplored opportunities for modeling human–agent and agent–agent misalignment. By synthesizing trends, challenges, and open research directions, this review provides a new roadmap for applying hypergame theory to enhance the realism and effectiveness of strategic modeling in dynamic multi-agent environments.

Keywords: Theory of Mind, Multi-Agent Systems, Game Theory, Hypergame Theory

1 Introduction

Strategic decision-making is a key issue in many real-world scenarios involving interactions among multiple agents, and as such is a fundamental agentic capability. Since agents in these settings must choose actions based on the actions of others, game theory naturally emerges as the discipline providing formal tools to model, design, and analyze such systems (Rasmusen, 1990). However, real-world environments are often marked by uncertainty, misaligned perceptions, and nested beliefs, which frequently render traditional models inadequate. To address these limitations, researchers have developed frameworks that incorporate cognitive constraints, recursive reasoning, and heterogeneous agents exhibiting bounded rationality (Harsanyi, 1967-1968; Simon, 1972; Nagel, 1995; Camerer et al., 2004; Dekel and Siniscalchi, 2015).

Simpler and less expressive models, such as the k-level theory and the cognitive hierarchy model (Camerer et al., 2004), acknowledge differences in players' reasoning capacities but offer limited insight into the structure and underlying nature of these differences. Epistemic game theory provides a formal language for representing player types, belief hierarchies, and uncertainty; however, its predominantly probabilistic models only partially account for the mechanisms that give rise to these cognitive variations (Dekel and Siniscalchi, 2015).

In contrast, hypergame theory advances the game-theoretic paradigm by explicitly modeling subjective representations of the strategic scenario – so-called *perceptual games* – in which agents may disagree about the components of the game: the participating players, the available actions, or the payoff structure (Bennett, 1977). This capability makes hypergames particularly suitable for multi-agent system (MAS) frameworks that aim to emulate human-like reasoning and cognitive sophistication. Such suitability aligns with recent interest in theory of mind (ToM) – a psychological model of human reasoning that characterizes the ability to attribute mental states to oneself and others (Premack and Woodruff, 1978) – and its integration into agentic artificial intelligence (Rocha et al., 2023).

Originally developed as an analytical framework for post-hoc conflict analysis, hypergame theory was not designed for dynamic or interactive agent-based systems. As a result, despite its descriptive power and emerging practical applications, its full potential in MAS environments remains underutilized. This systematic review aims to address that gap by identifying existing agent-compatible applications of hypergame theory, analyzing patterns of practical integration, and highlighting both structural challenges and emerging opportunities at the intersection of hypergames and multi-agent systems.

We make several contributions. First, we provide a formal introduction to hypergame theory, including its foundational concepts and its two major revisions: hierarchical hypergames and the hypergame normal form (HNF). Building on this foundation, we define agent-compatibility criteria for hypergame models and apply these criteria in a systematic review of 44 selected papers from the hypergame-theoretic literature. We develop an agent-based classification framework to examine how these works integrate hypergame-theoretic models, categorizing them by domain and subdomain, practical applicability, and degree of integration – from conceptual inspiration through partial implementation to full deployment. We also construct a structured

taxonomy of hypergame formalisms, distinguishing multi-level and HNF models, and map their usage across planning, reasoning, learning, and uncertainty modeling tasks. Furthermore, we synthesize trends and usage patterns, showing that while theoretical integrations of hypergames remain prevalent, practical deployments often rely on simplified or flattened models. Finally, we discuss gaps in infrastructure, such as the lack of agent-oriented hypergame modeling languages, and suggest avenues to address these gaps by drawing inspiration from formal languages like epistemic logic and GDL-III, which are already used in MAS research.

The remainder of this paper is organized as follows. Section 2 reviews the foundations of hypergame theory, including its game-theoretic roots, the formalization of games, and the development of hierarchical and HNF-based hypergame models. Section 3 presents the methodology and results of our systematic review, covering publication trends, a taxonomy of hypergame models, and an analysis of their application in practice – including application domains, integration methodologies, fidelity, and task alignment. Section 4 discusses the survey results, with particular attention to usage patterns, domain-model associations, and the need for agent-based hypergame languages. Section 5 situates this work within the broader literature, and Section 6 concludes with a summary of findings and directions for future research.

2 Hypergame Theory

In this section, we first establish the foundations of game-theoretic reasoning by introducing classical models of strategic interaction. We then introduce the theory of hypergames as an extension of game theory that relaxes the assumptions of common knowledge and shared perceptions. Finally, we present two major formalisms of hypergame theory: the Hypergame Normal Form (HNF) and Hierarchical Hypergames, both of which have served as theoretical bases for modeling misaligned perceptions and recursive reasoning in multi-agent systems. For the purpose of introducing the relevant concepts, we use the two-player rock-paper-scissors game as a running example – a well-known, simple setting where both players must simultaneously choose one of three available actions: Rock, Paper, or Scissors.

2.1 Game-theoretic Foundations

While decision theory provides tools to reason about how an individual agent can derive choices, when another decision-maker is involved the player may have to explicitly reason about how the opponent is trying to solve the same problem. Game theory is better equipped for such analyses, as it deals with conflicts where decisions and their outcomes are interdependent between agents (Myerson, 1984).

2.1.1 Elements of a Game

(Rasmusen, 1990) introduces the key elements of game-theoretic interactions through the acronym *PAPI*: Players, Actions, Payoff and Information. We define these four elements as follows:

Players

Players are the decision-making entities involved in the strategic interaction. Game-theoretic models typically assume that players are (Myerson, 1984):

- **Rational:** they consistently act to maximize their own utility functions;
- **Strategically intelligent:** they understand the structure of the game and anticipate that others behave similarly.

Actions

Actions represent the immediate choices available to each player in the game they are involved with. In contrast, strategies are agents' plans of actions based on the observed interaction history (Osborne and Rubinstein, 1994). We distinguish between two classes of strategies (Luce and Raiffa, 1989):

- **Pure:** deterministic mappings from states or histories to actions;
- **Mixed:** probability distributions over pure strategies.

Payoffs

Payoffs quantify the preferences of each player over possible outcomes. They are typically represented as real-valued utility functions $U_i : A \to \mathbb{R}$, where A is the joint action space. Numerical payoffs enable the use of optimization and equilibrium analysis (von Neumann and Morgenstern, 1944; Gibbons, 1992).

Information

Information concerns players' knowledge at the time of making a decision. This knowledge encompasses a player's image of the other players, the options available to everyone, and the historical moves. We classify games based on three information criteria:

- **Perfect vs. Imperfect Information**: Whether agents can fully observe others' actions (Osborne and Rubinstein, 1994);
- **Complete vs. Incomplete Information**: Whether agents can fully observe others' payoffs (Gibbons, 1992);
- **Symmetric vs. Asymmetric Information**: Whether agents can access the same quality and quantity of information (Rasmusen, 1990).

2.1.2 Formalizing Games

A game-theoretic model of rock-paper-scissors can be formally defined as a tuple $G = (N, A, u)$ where:

- $N = \{1, 2\}$ is the set of players
- $A = A_1 \times A_2$ is the set of joint actions, where for each player $i \in N$:
  - $A_i = \{\text{Rock}, \text{Paper}, \text{Scissors}\}$ is the action set of player i
  - A pure strategy for player i is an element $a_i \in A_i$
  - A mixed strategy for player i is a probability distribution σ_i over A_i

- $u = (u_1, u_2)$ are the utility functions where for each player $i \in N$:

$$u_i : A \to \mathbb{R} \text{ defined as:}$$

$$u_i(a_i, a_j) = \begin{cases} 1 & \text{if } (a_i, a_j) \in W_i \\ 0 & \text{if } a_i = a_j \\ -1 & \text{if } (a_i, a_j) \in L_i \end{cases}$$

  where:

  - $W_i = \{(\text{Rock}, \text{Scissors}), (\text{Paper}, \text{Rock}), (\text{Scissors}, \text{Paper})\}$ are winning outcomes
  - $L_i = \{(\text{Rock}, \text{Paper}), (\text{Paper}, \text{Scissors}), (\text{Scissors}, \text{Rock})\}$ are losing outcomes

Expected Utility

When players employ mixed strategies – probability distributions over their pure actions – the outcome of the game becomes probabilistic. The expected utility for a player is then calculated as the average payoff, weighted by the likelihood of each joint action occurring under the players' chosen strategies (von Neumann and Morgenstern, 1944). Formally, given mixed strategies σ_i and σ_j for players i and j, the expected utility for player i is:

$$EU_i(\sigma_i, \sigma_j) = \sum_{a_i \in A_i} \sum_{a_j \in A_j} \sigma_i(a_i)\sigma_j(a_j)u_i(a_i, a_j)$$

Rational players are typically assumed to choose strategies that maximize their expected utility, forming the basis of solution concepts such as the mixed-strategy Nash equilibrium.

Nash Equilibrium

The Nash Equilibrium (NE) is a fundamental solution concept in game theory, capturing the state when neither player can improve their payoff by a unilateral change of strategy (Nash, 1950). Formally, a strategy profile $a^* = (a_i^*, a_{-i}^*)$ – with $-i$ denoting the opponent(s) – is a NE, for each player $i \in N$:

$$u_i(a_i^*, a_{-i}^*) \geqslant u_i(a_i, a_{-i}^*) \quad \forall a_i \in A_i$$

In case of mixed strategies, the Mixed Strategy Nash Eqilibrium (MSNE) is defined analogously: each player's mixed strategy must be a best response to the mixed strategies of others. Formally:

$$EU_i(\sigma_i^*, \sigma_{-i}^*) \geqslant EU_i(\sigma_i, \sigma_{-i}^*) \quad \forall \sigma_i \in \Delta(A_i),$$

where $\Delta(A_i)$ denotes the set of mixed strategies over A_i.

In the case of rock-paper-scissors, no pure strategy profile constitutes a Nash equilibrium. However, a unique mixed strategy Nash equilibrium exists where each player chooses Rock, Paper, and Scissors with equal probability (Batzilis et al., 2019).

2.2 Games of Misperception

Real-world agents may differ in their abilities to observe and process information, potentially leading to distinct internal models of others and misaligned understandings of the conflict, the available actions, or the associated outcomes. Hypergames provide an explicit model of such perceptions by capturing each agent's subjective interpretation of the situation (Bennett, 1980). Furthermore, hypergame analyses rely on preference relations rather than numerical payoffs, resulting in a qualitative model of desirability. This approach aims to mitigate challenges associated with assigning precise numerical values to real-world outcomes.

Hypergames are considered models of bounded rationality (Simon, 1972; Vane, 2006) because they support reasoning in which an agent may perceive an opponent's choice as irrational, even though that choice is rational from the opponent's own perspective (Kovach et al., 2015). Formally, a hypergame is composed of each player's *perceptual game* – a version of the conflict that represents the player's interpretation – and is defined as the tuple $H = (N, \{G_i\}_{i \in N})$ where:

- $N = \{1, 2, \ldots, n\}$ is the set of agents;
- For each agent $i \in N$, G_i represents agent i's perceived game, where $G_i = (N_i, A_i, R_i)$ with:
  - $N_i \subseteq N$ is the set of agents as perceived by i;
  - $A_i = \times_{j \in N_i} A_{ij}$ is the joint action space as perceived by i, where:
    - A_{ij} is i's perception of j's available actions;
  - $R_i = \{R_{ij}\}_{j \in N_i}$ is the set of preference relations as perceived by i, where:
    - $R_{ij} \subseteq A_i \times A_j$ is i's perception of j's preference relation;
    - For outcomes $x, y \in A_i$, $x > y$ indicates that i prefers outcome x to outcome y.

Note: elements with indices ii denote player $i's$ own interpretation of the game's components.

To illustrate misaligned perceptions, consider an asymmetric version of the rock-paper-scissors game modeled as a hypergame in which one agent's action space is constrained. Let $H = (N, \{G_1, G_2\})$, where:

- G_2 is Agent 1's perceived game:
  - $N_1 = \{1, 2\}$;
  - $A_{11} = A_{12} = \{\text{Rock}, \text{Paper}, \text{Scissors}\}$;
  - $R_{11} = R_{12} =$

$$\{(P, R), (S, P), (R, S)\} > \{(R, R), (P, P), (S, S)\} > \{(S, R), (R, P), (P, S)\};$$

- G_2 is Agent 2's perceived game:
  - $N_2 = \{1, 2\}$;
  - $A_{21} = A_{22} = \{\text{Rock}, \text{Paper}\}$;

– $R_{21} = R_{22} =$

$$\{(R,R),(P,P)\} > \{(R,P),(P,R)\};$$

This example captures a situation where Scissors is not recognized as an available action by Agent 2, whereas Agent 1 perceives the complete action space. This form of perceptual asymmetry mirrors the classic "Fall of France" hypergame analysis (Bennett and Dando, 1979). The resulting discrepancy in perceived game structure can lead to unexpected outcomes – referred to in hypergame theory as *strategic surprise* (Bennett, 1980).

2.3 Hierarchical Hypergames

Hypergame theory introduces a framework for modeling individual agents' subjective interpretations of the conflict they are involved in. By incorporating higher-order perceptions (Bennett, 1980), hypergames enable the formalization of recursive reasoning – e.g., a decision-maker r might consider what q believes about how p perceives the situation. These are called m-th order expectations, where m corresponds to the depth of the perceptual chain. (Wang et al., 1988) later formalized and extended Bennet's framework into a hierarchical hypergame analysis structure, addressing the overly abstract nature of earlier definitions and proposing a more rigorous, standardized representation of multi-level perceptual games. In this extended formalism, a 0^{th}-level hypergame is equivalent to a standard game-theoretic without misaligned perceptions and is defined as a tuple $H^0 = (N, P, S, U, V)$, where:

- $N = \{i \mid 1 \leqslant i \leqslant n\}$ is the set of players;
- $P = \{P_i \mid i \in N\}$ is the set of players option sets, where:
  - P_i is the set of actions available to player i;
- $S = \{S_i \mid i \in N\}$ is the set of players' strategy sets, where:
  - $S_i \subseteq P_i$ is the set of actions player i considers for G;
- U is the set of all possible outcomes, where:
  - Each outcome $u \in U$ is a tuple of chosen strategies $u = (s_1, \ldots, s_n)$;
  - $U = \times_{i \in N} S_i$ represents all possible strategy combinations;
- $V = \{V_i \mid i \in N\}$ is the set of player preferences, where for each player i:
  - $V_i = (U, R_i)$ where R_i is a preference relation over U, or
  - $V_i = (U, u_i)$ where $u_i : U \to \mathbb{R}$ is a utility function.

The distinction between P_i and S_i allows the modeling of agents' beliefs about physical or cognitive constraints. A zero-level hypergame assumes shared perceptions – there are no subjective differences between agents' views of the game.

Higher-order Reasoning

To support higher-order reasoning, Wang et al. introduced the notion of *perceptual functions*. Each player's *viewpoint* encodes their holistic interpretation of the game, while individual perceptions capture their interpretation of specific components. For

any player i, a perception is defined as a mapping:

$$f_i : \Gamma_i \to \Gamma_{ij},$$

where:

- Γ_j is a component from player j's game (from $\{N, P, S, U, V\}$);
- Γ_{ij} denotes player i's internal model of Γ_j;
- For each element $\gamma \in \Gamma_j$, $\varphi = f_i(\gamma)$ is i's perceived *image* of γ.

In the case of nested beliefs, player i's interpretation is composed via the product mapping:

$$f_i \circ f_j : \Gamma_k \to \Gamma_{ijk}.$$

The resulting image is the product of individual perceptual functions:

$$\varphi = f_i(f_j(\gamma)) = f_i \circ f_j(\gamma) = f_{ij}(\gamma).$$

Misperceptions can arise across all game components when the image $\varphi = f_i(\gamma)$ deviates from the true $\gamma \in \Gamma_j$. A perception is:

- Underperceived if it omits elements present in Γ_j;
- Overperceived if it includes elements not present in Γ_j;
- Otherwise, a misperception if $\varphi \neq \gamma$.

Higher-order Games

In cases where players develop misaligned perceptions and expectations about others' games, we define higher-order hypergames as follows:

- A **first-level hypergame** assumes that players are unaware of any perceptual misalignment. Each player i' viewpoint is captured by their own perceptual game G_i, forming the composite hypergame $H^1 = \{G_i \mid 1 \leqslant i \leqslant n\}$. Each perceptual game is defined using potentially distinct sets N_i, P_i, S_i, U_i, V_i.
- A **second-level hypergame** arises when at least one player is aware that others may have different viewpoints. $H^2 = \{H_i^1 \mid 1 \leqslant i \leqslant n\}$, where each H_i^1 now contains the set of perceptual games $G_{ij} = (N_{ij}, P_{ij}, S_{ij}, U_{ij}, V_{ij})$ that player i believes player j uses: $H_i^1 = \{G_{ij} \mid 1 \leqslant j \leqslant m\}$.
- A **third-level hypergame** extends this recursion once more: a player i models what player j believes about what player k' game. We have $H^3 = \{H_i^2 \mid 1 \leqslant i \leqslant n\}$, where each individual game is composed of first-level hypergames $H_i^2 = \{H_{ij}^1 \mid 1 \leqslant j \leqslant m\}$ and each first-level hypergame is a set of standard games $H_{ij}^1 = \{G_{ijk} \mid 1 \leqslant k \leqslant h\}$, where each percpetual game is $G_{ijk} = (N_{ijk}, P_{ijk}, S_{ijk}, U_{ijk}, V_{ijk})$.

At each level, the formal structure is extended by an additional layer of subjective (hyper)games built on nested perceptions. While the mathematical framework supports L^{th}-order hypergames of arbitrary depth, we restrict the discussion to level 3, reflecting empirical observations that human strategic reasoning seldom exceeds three levels (Camerer et al., 2004) and balancing expressiveness and computational tractability.

Hypergame Nash Equilibrium

The Hypergame Nash Equilibrium (HNE) extends the classical Nash Equilibrium to settings where players may have divergent perceptions of the game (Wang et al., 1988). The HNE formalizes mutual best responses in this context of misaligned perceptions; that is, the HNE is an NE in each player's corresponding subjective game. Given a strategy profile $a^* = (a_i^*, a_{-i}^*)$:

- In a first-level hypergame $H^1 = \{G_i \mid 1 \leqslant i \leqslant n\}$, a^* is a HNE if, for each player $i \in N$, a^* is a NE in their subjective game G_i;
- In a second-level hypergame $H^2 = \{H_i^1 \mid 1 \leqslant i \leqslant n\}$, a^* is a HNE if, for each player $i \in N$ and $j \in N_i$, a^* is a NE in G_{ij};
- In a third-level hypergame $H^3 = \{H_i^2 \mid 1 \leqslant i \leqslant n\}$, a^* is a HNE if, for each player $i \in N$, $j \in N_i$ and $k \in N_{ij}$, a^* is a NE in G_{ijk}.

2.4 Hypergame Normal Form

The normal form is a foundational representation in game theory used to express strategic interactions among decision-makers, where players' actions and the corresponding payoffs are captured in a table, typically referred to as the payoff matrix (Rasmusen, 1990). The Hypergame Normal Form (HNF) extends the standard normal form to capture a single player's beliefs – drawing on decision-theoretic concepts (Hansson, 2011) – about their opponent's reasoning (Vane and Lehner, 2000). Originally proposed for analyzing complex military planning scenarios, HNF enables an agent to reason over multiple possible opponent strategies, each corresponding to a different perceptual subgame. Table 1 illustrates an HNF matrix for the rock-paper-scissors game used in previous examples from the row player's perspective. The HNF consists of:

- The **Full Game** (bottom right quadrant) represents the "expert's view" – a standard $n \times m$ game-theoretic payoff matrix, where n and m denote the set of actions available to the row and column players, respectively;
- The **Column Mixed Strategies** (top right quadrant) or CMS represent possible opponent strategies, including:
  - $\mathbf{C_0}$: the Nash equilibrium mixed strategy (NEMS);
  - $(\mathbf{C_1}, \ldots, \mathbf{C_k})$: the row player's belief-induced subgames representing alternative opponent mixed strategies over $(S_1, \ldots, S_m)$;
  - $\mathbf{C_\Sigma}$: a weighted summary of beliefs, computed as:

$$S_j = \sum_{k=0}^{K-1} P_k c_{kj}, \quad \text{for } j = 0, 1, \ldots, n$$

    where P_k is the weight for belief context k and c_{kj} is the weight for strategy j in context k.

- The **Row's Belief Context** (upper left quadrant) contains probabilities assigned to each opponent CMS. These are analogous to perceptual functions in multi-level

hypergames but focused on strategy selection rather than game structure. They capture the player's beliefs about the likelihood of each opponent's reasoning mode, that is, the probability that the corresponding subgame (CMS row) is active;

- The **Row Mixed Strategy** (RMS) zone (bottom left quadrant) captures the row player's responses to each belief context, resembling second-order beliefs in multi-level hypergames. RMS entries are not prescriptive but inform planning.

					Column Mixed Strategy		
Row's Belief Context				C_Σ	S_1	S_2	S_3
P_2	0.5			C_2	$\frac{1}{2}$	$\frac{1}{2}$	—
P_1		0.25		C_1	0.326	0.3456	0.3284
P_0			0.25	C_0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
H_1	R_2	R_1	R_0		**Rock**	**Paper**	**Scissors**
H_{11}	0.2	0.3284	$\frac{1}{3}$	**Rock**	0	-1	1
H_{12}	0.6	0.326	$\frac{1}{3}$	**Paper**	1	0	-1
H_{13}	0.2	0.3456	$\frac{1}{3}$	**Scissors**	-1	1	0
Row Mixed Strategy					**Full Game**		

Table 1: Hypergame Normal Form matrix for rock-paper-scissors comprising of: (1) Row's Belief Context with probabilities P_k for each belief context, (2) Column Mixed Strategies showing opponent mixed strategies C_k and their aggregate C_Σ, (3) Row Mixed Strategy containing R_k responses to each belief context, and (4) Full Game with cardinal utilities. Hyperstrategies H_{ij} represent evaluated plans, expressed as probability distributions over row strategies. Each C_k and R_k represents a mixed strategy vector summing to 1.

The game in Table 1 defines the row player's beliefs about the column player in rock-paper-scissors. C_0 corresponds to the mixed strategy Nash equilibrium (uniform randomization), while C_1 reflects empirical distributions derived from online gameplay data (Batzilis et al., 2019). In C_2, the row player believes that the opponent cannot play Scissors. The belief context assigns the highest probability to C_2, indicating that the row player expects this restricted subgame to be most likely. The RMS entries mirror responses to each C_k. Additionally, the HNF matrix can include expected utility (EU) and hypergame expected utility (HEU) calculations as additional rows beneath the matrix – omitted in Table 1 for clarity.

Hyperstrategies

Hyperstrategies are mixed strategies over RMSs, represented as vectors of assigned probabilities adjacent to the RMS quadrant. We include a hypothetical hyperstrategy H_1 in Table 1 for illustration. Vane (Vane and Lehner, 2000) defines four

hyperstrategies that meet the following efficiency criterion:

$$\text{Eff}(hs) \leftrightarrow \text{EU}(hs, C_\Sigma) \geqslant \text{EU}(\text{NEMS}(m \times n)).$$

These are defined as follows:

- **NEMS**: NEMS is the baseline Nash equilibrium mixed strategy derived from the full game;
- **Modeling Opponent (MO)**: MO is a decision-theoretic strategy selecting the best response based on CMS and RMS utilities;
- **Pick Subgame (PS)**: PS selects the Nash equilibrium of the subgame identified by MO;
- **Weighted Subgame (WS)**: WS is a hybrid strategy that chooses the best option based on the combination of the MO-specified CMS and R_0 weighted by the corresponding belief context probabilities.

Hypergame Expected Utility

Vane (Vane, 2000) introduced hypergame expected utility (HEU) as a bridge between decision theory and game theory. HEU incorporates a parameter g – the "fear of being outguessed" – and the worst-case outcome G, defined as the lowest expected utility for the row player. Then, given a hyperstrategy hs, we define the HEU as:

$$\text{HEU}(hs, g) = (1 - g) \cdot \text{EU}(hs, C_\Sigma) + g \cdot \text{EU}(hs, G).$$

The relative performance of hyperstrategies such as MO, PS, and WS varies with the uncertainty parameter g, with NEMS becoming increasingly dominant as uncertainty rises. The analysis of g's effects – detailed in (Vane, 2000) – suggests that when agents hold accurate beliefs about their opponent's strategy, HEU-optimized hyperstrategies can outperform standard equilibrium analysis.

3 Hypergame Theory in Multi-Agent Systems

While hypergame theory is a model of bounded rationality that can provide multi-agent systems with a framework that can model agent heterogeneity and environmental uncertainties to a greater level of detail than probabilistic methods, it was originally designed to be an analytical framework targeting post-hoc studies of complex conflicts instead of providing tools for the decision-maker on the spot. While hypergame models have yet to become prevalent in dynamic settings multi-agent systems entail, they have a potential to overcome the constraints posed by typical game-theoretic models on the complexity agent-based frameworks can capture (Sasaki, 2021). Motivated by this identified potential, in this section we provide a systematic review of hypergame-theoretic works focusing on studies that entail agent-based applications or the potential to be integrated into dynamic agentic systems.

3.1 Methodology

Our systematic survey of hypergame theory in the context of multi-agent systems follows a multi-step screening and classification process. Our goal was to identify works where hypergame-theoretic models are not only referenced, but explicitly applied in ways that support agentic reasoning, decision-making, or strategic interaction within MAS-compatible contexts. Figure 1 provides an overview of this selection funnel. Starting with an inclusive set of keyword-matched results, we progressively filtered out duplicates, non-English works, irrelevant uses, and purely analytical studies lacking actionable agent-level integration. The steps are outlined below:

Initial Collection

We conducted a keyword-based query based on Google Scholar using the term "hypergame theory", which yielded 320 unique results as of June 2025.

Duplicate Removal

We removed 30 entries due to duplicity. Removed entries included different versions of the same manuscript (e.g., preprints and published versions) as well as overlapping content such as thesis chapters and standalone publications.

Language Filter

We excluded 17 papers for being written in languages other than English.

Survey and Meta-reviews

We identified 6 publications that were either surveys of hypergame theory or broader reviews that discussed hypergames as part of a more general framework. These were retained for comparative purposes but not included in the core set analyzed for model classification.

Relevance Filtering

While hypergame theory is a distinct subfield within game theory, the term is occasionally used in unrelated contexts or cited without substantive application. To ensure relevance, we manually examined each entry's abstract, methods, references, and in-text keyword usage. We excluded 154 papers that either lacked an actual implementation of hypergames, used the concept only metaphorically, or mentioned it in passing without integration into the modeling framework.

Agent-compatibility Filter

From the remaining 113 works, we applied a final inclusion criterion focused on agent compatibility. Specifically, we retained only those works in which hypergames were used to support or influence agent behavior, strategic decision-making, or reasoning under uncertainty – either in simulated, practical, or theoretical settings with agent-level implications. This step excluded 69 additional papers that employed hypergames solely from an analyst perspective, or in a way that did not feed back into the agent's own actions or perceptions.

As a result, we identified a final corpus of 44 papers featuring agent-compatible hypergame-theoretic models. Although only three of these explicitly formalize the agent architecture, the majority assume rational decision-makers with embedded utilities or subjective games. Due to the inconsistency in architectural specification, the focus of our review excludes implicit and explicit agent design.

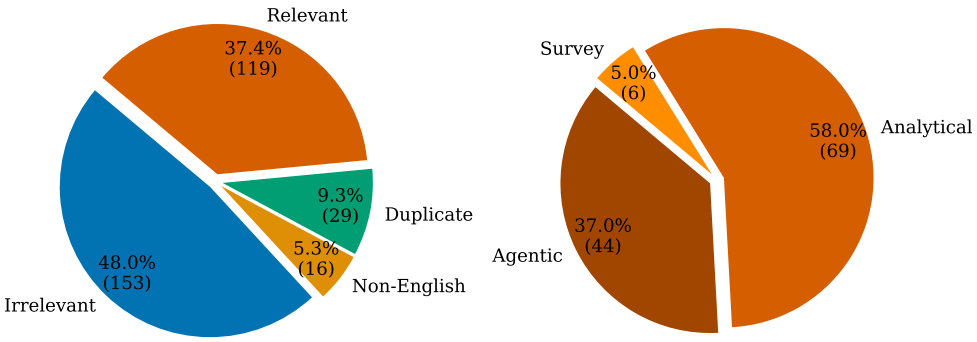

Fig. 1: Distribution of the evaluated 320 papers based on our selection criteria. The left chart illustrates the composition of papers eliminated due to irrelevance, duplication, or being written in a language other than English. The chart to the right illustrates the agent-compatibility evaluation that resulted in the final 44 papers considered for our review.

3.2 Hypergame-theoretic Publication Trends

Bennett introduced the theory of hypergames as an analytical framework extending standard game-theoretic frameworks, emphasizing the modeling ability to account for perceptual misalignments. Vane's hypergame normal form revised Bennett's hypergame representation, providing an extended decision-theoretic matrix form focusing on one player's belief systems and reasoning. Early applications of both frameworks solely entailed post-hoc or hypothetical analyses, with no demonstrations on how a decision-maker can benefit from hypergame theory in real-time. While both original authors repeatedly highlighted the potential of hypergames to be used in dynamic agent-based applications and both made attempts of "live" hypergame analyses, the adoption of hypergame theory, in multi-agent literature in particular, remains to take off. Nonetheless, our review reveals a trend of growing number of MAS-related

hypergame-theoretic publications in recent years. We see a shift towards agent-based applications, as illustrated on Figure 2. Interestingly, (Kovach et al., 2015) stated that

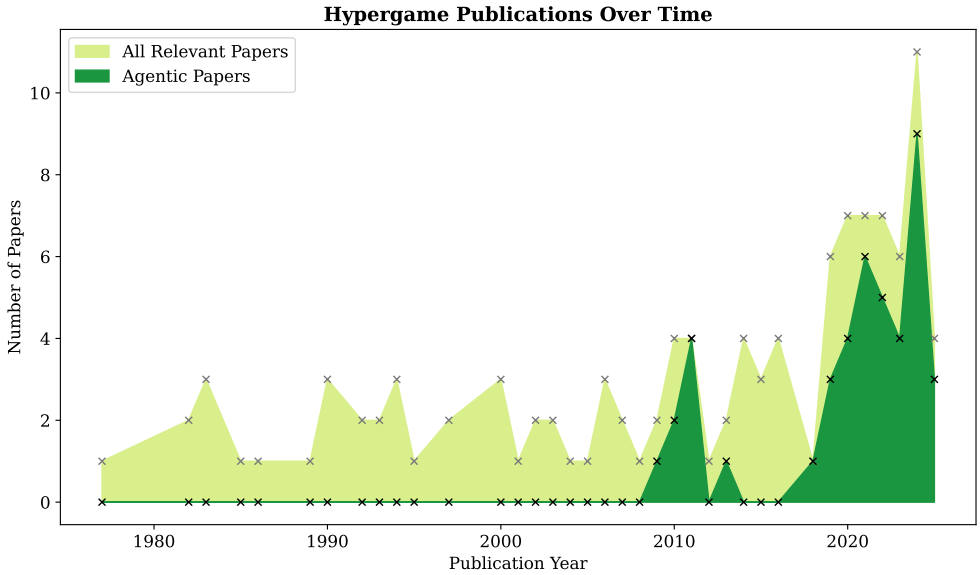

Fig. 2: Frequency of all hypergame-relevant and agentic publications over time.

"Unfortunately very little in the way of hypergame theory application has been done in the arena of cyber warfare", while a decade later the current shift is dominantly lead by attack-defense modeling and simulations in cybersecurity – as demonstrated by Figures 6, 7, 9 and the various cybersecurity-related surveys covering hypergame-based approaches we discuss in section 5.

Hypergame theory is a niche field that has not yet gained traction, with a modest number of publications, dominated by a historical concentration of contributions from the original authors and their collaborators. To investigate whether recent trends – such as increasing domain diversity and methodological variation – are accompanied by a broader dissemination of authorship, we conducted an analysis of author frequencies and co-authorship patterns across the 44 papers reviewed. We extracted all named authors from each publication, accounting for different name formats and removing placeholder or malformed entries. First, we identified individual authors associated with at least three of the reviewed papers.

Then, to identify collaborative groups, we constructed a co-authorship graph in which nodes represent individual authors and edges indicate co-authorship on at least one paper. Groups were then defined using connected components of this graph, with each component representing a set of authors linked by direct or indirect collaboration. For each group, we computed the number of papers coauthored by at least two members, thereby capturing the extent and productivity of distinct author clusters.

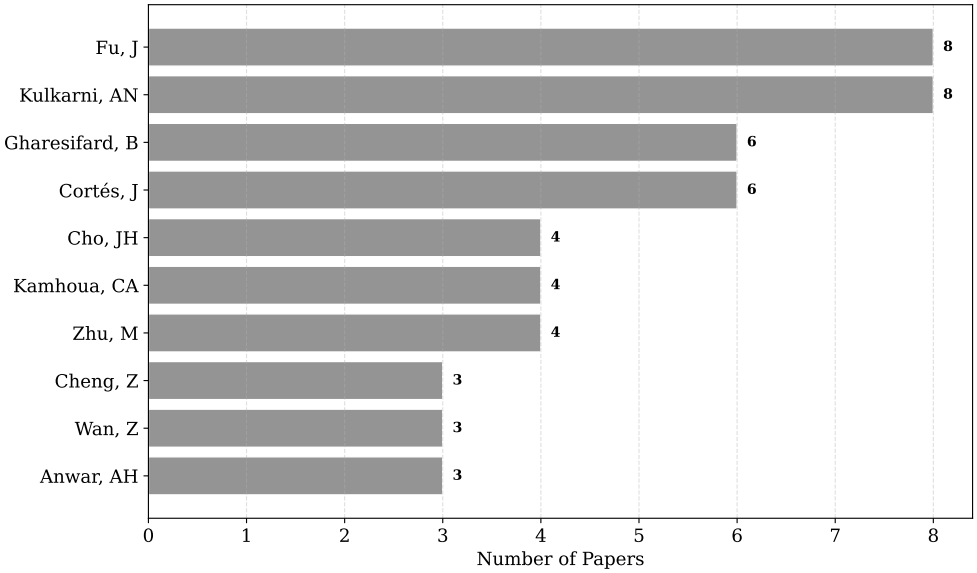

Fig. 3: Publication counts per authors with at least 3 associated works.

Table 2: Identified coauthor groups with at least 3 shared papers. Member names are shown by surname only.

Group	Shared Papers	Members
G5	11	Cohen, Fried, Fu, Gaglione, Han, Hemida, Kamhoua, Kulkarni, Leslie, Li, Liu, Luo, Ma, Shi, Udupa, Xi
G6	6	Cortés, Gharesifard
G1	4	Anwar, Cho, Kamhoua, Singh, Wan, Zhu
G8	3	Chen, Cheng, Gao, Guan, He, Hong, Ma, Yuan

As shown in Figure 3, ten authors are associated with at least three papers, with nearly a third of reviewed papers attributable to two recurring author pairs: Kulkarni & Fu and Gharesifard & Cortés. Furthermore, as captured by Table 2, we found four groups of collaborators with at least three relevant publications, where Group 5 – Kulkarni & Fu et al. – is responsible for a quarter of publications. At the same time, over half of the papers originate from authors or groups with fewer than three publications, suggesting a shift from concentrated research programs toward more diverse, application-specific uses of hypergames, potentially reflecting growing interest and wider adoption of the theory across domains – a finding well-aligned with the publication trends identified via our classification framework and Figure 2.

3.3 Taxonomy of Hypergame Models

In this section, we provide a detailed taxonomy of hypergame models identified across the surveyed papers. We categorize the models according to three key dimensions: the

foundational hypergame concept and game dynamics, formalizations of hypergame frameworks, and the computational tasks hypergame-theoretic approaches address. This taxonomy facilitates a structured understanding of how different hypergame implementations vary in their theoretical assumptions, representational complexity, and practical application scenarios, enabling researchers to select or develop hypergame models tailored to specific multi-agent system requirements.

Game Dynamics

First, we looked at whether the hypergame models entail a single round of interaction or repeated, iterative decision-making. The distribution across hypergame concepts and game dynamics is characterized by 30 applications that entail repeated interactions, while 14 models rely on one-shot games. This suggests a shift in applications trends, as historically works primarily involved one-shot post-hoc studies.

Hypergame Concept

We also distinguished works by whether they were inspired by the original (Bennett, 1980) or extended (Wang et al., 1988) hypergame framework – collectively denoted as **Multi-level Hypergames (MLH)** – or based on the **Hypergame Normal Form (HNF)**. We found 7 papers that refer to HNF, 35 rely on Multi-level Hypergames, and 2 papers use components from both concepts, which we denote as **Hybrid** approaches.

Hypergame Formalizations

Next, we identified common practical hypergame model definitions, which we classify as follows:

- **HNF-based** solutions populate complete Hypergame Normal Form (HNF) matrices, which agents directly utilize for decision-making.
- **HEU-based** methods utilize the Hypergame Expected Utility derived from HNF to help agents select optimal strategies.
- **L-th Order Multi-level Hypergames** populate hierarchical hypergame structures, enabling agents to form nested beliefs.
- **Flattened L-th Order Multi-level Hypergames** employ subjective games but exclude the hierarchical tree structure.
- **Bayesian L-th Order Hypergames** implement hybrid belief models in which hypergames represent perceptions, but agents perform decision-making using Bayesian reasoning.
- **Perceptual Game-based** implementations enable agents to develop hypergame-like beliefs without defining a complete multi-level hypergame structure.
- **HNE-based** frameworks involve decision-makers explicitly relying on the Hypergame Nash Equilibrium concept.
- **Graph-based** methods incorporate spatial hypergames and strategies utilizing graph-based representations.
- **Stackelberg Hypergames** extend the leader-follower mechanism of Stackelberg games by integrating hypergame-theoretic subjective perceptions.

The distribution of model usage across the surveyed papers is depicted in Figure 4: the majority of works implement a model based on the typical multi-level hypergame concept (L-th Order, Flattened L-th Order and Bayesian L-th Order) or hypergames on graphs. 5 works entail hybrid solution, where in addition to populating a hypergame model a hypergame-theoretic solution concept (HEU or HNE) is also implemented. In two cases, the hybrid approach entailed cross-framework implementations: multi-level-hypergame-based models integrating the hypergame expected utility concept from HNF.

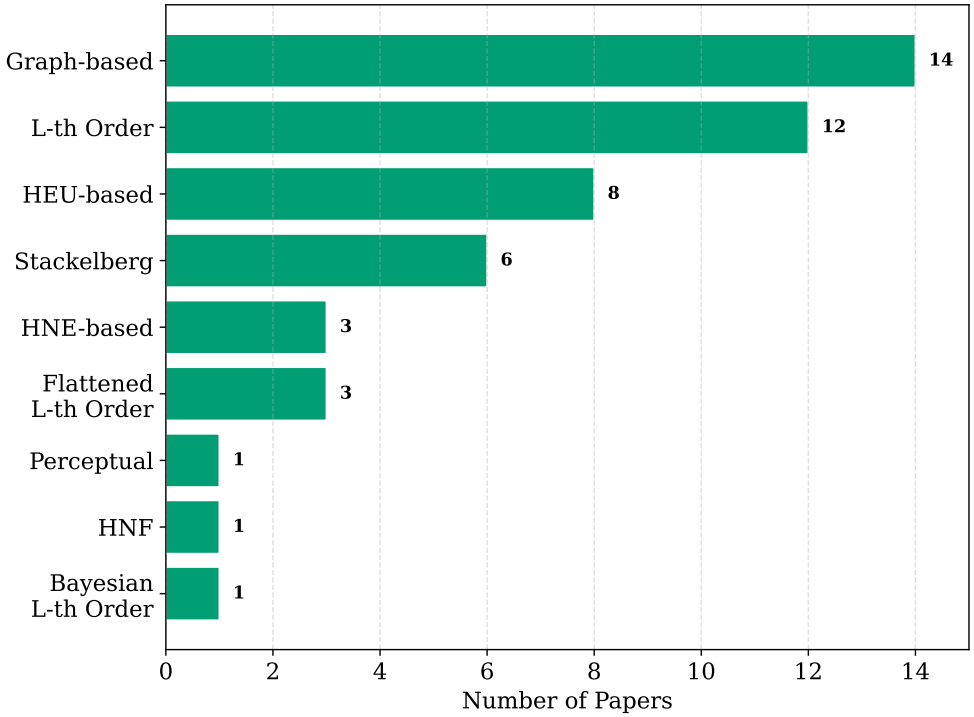

Fig. 4: Distribution of model types. Hybrid models that span multiple categories are counted in each relevant type.

3.4 Hypergames in Practice

Hypergames, despite being originally developed as a post-hoc analytical framework, provide a promising tool for modeling realistic interaction scenarios with misaligned perceptions in multi-agent systems. Practical implementations employing, to varying extents, hypergame concepts emerge in such diverse fields as cybersecurity, robotics, and social simulations. To identify current patterns in the practical implementation of hypergames, as well as the challenges and open research problems, we

analyze the papers that implement hypergame concepts in terms of the application domain (Section 3.4.1), hypergame integration methodology (Section 3.4.2), integration fidelity (Section 3.4.3), and finally, the role of the hypergame model within the given framework (Section 3.4.4).

3.4.1 Application Domains

To investigate how hypergame theory has been adopted across different areas of research and application, we classified the 44 agent-compatible papers according to their primary domain of use. Each paper was assigned to one of six high-level domains, which reflect the thematic focus of the study: *Game Theory*, *Cybersecurity*, *Social Simulations*, *Robotics*, *Multi-agent Systems*, and *Communications*. Within each domain, we identified recurring sub-domains that capture specific use cases, each briefly introduced in the following subsections and highlighted in **bold** to aid navigation. In Figure 5, we provide an overview of the reviewed works across the identified domains and sub-domains, which are summarized as follows.

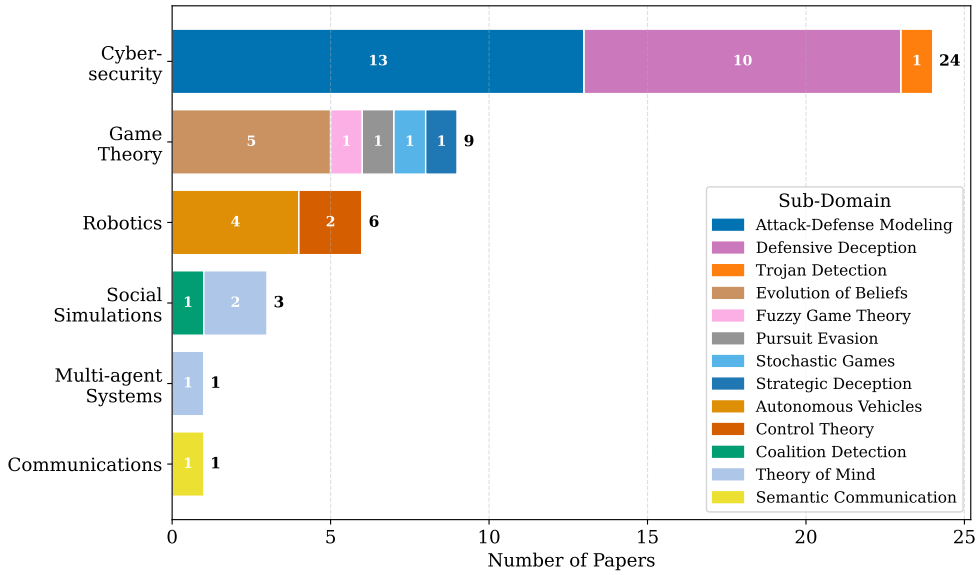

Fig. 5: Distribution of agent-compatible hypergame papers across high-level application domains and their associated sub-domains, where sum-domains can overlap between application domains.

Game Theory Works that advance game theory in general, without applying to any specific practical domain, are categorized under the general domain of *Game Theory*. These papers utilize hypergame-theoretic tools to define equilibrium concepts, model learning dynamics under misperception, or devise strategic responses to asymmetry

and deception. For example, (Song et al., 2009) integrates **Fuzzy Game Theory** with hypergames by incorporating fuzzy strategies and imprecise preference perceptions. In contrast, (Yuan et al., 2024) focuses on **Deception** and uses hypergames to analyze how misinformation affects equilibrium strategies. Additionally, (Li et al., 2024) proposes a hypergame-based Q-learning approach to solve **Pursuit-Evasion** games, while (Shi et al., 2023) develops a hypergame-theoretic model of deception in concurrent **Stochastic Games**. Lastly, various works leverage the hypergame-theoretic **Evolution of Beliefs** in adversarial games characterized by misperceptions. These studies define mechanisms for belief updates (Gharesifard and Cortés, 2010; Gharesifard and Cortes, 2012) that target the learning of equilibria (Gharesifard and Cortes, 2011a; Gharesifard and Cortés, 2011b) and belief manipulation (Gharesifard and Cortés, 2011c).

Cybersecurity Cybersecurity represents the most active domain for agent-compatible hypergame applications. This prevalence can be attributed to the inherent suitability of hypergames for capturing misinformation and modeling active deception – both critical issues within cybersecurity contexts. The majority of the surveyed works address **Attack-Defense Modeling** in general, formulating hypergames on graphs and applying linear temporal logic to guide deceptive strategy synthesis (Kulkarni and Fu, 2021a,b; Kulkarni et al., 2021; Li et al., 2023). Other examples include belief manipulation (Gharesifard and Cortés, 2014; Bakker et al., 2020), mission impact assessments employing hypergame-theoretic frameworks (Thukkaraju et al., 2023; Yoon et al., 2025), and other specialized hypergame-based methodologies for attack-defense modeling (House and Cybenko, 2010; Gutierrez et al., 2018; Bakker et al., 2021; Cheng et al., 2022). In addition, several works target a narrower area of attack-defense modeling. **Defensive Deception** entails the use of deceptive strategies to mislead the attacker, often focusing on optimizing strategic honeypot, decoy, and false target assignment (Xi and Kamhoua, 2020; Kulkarni et al., 2020; Wan et al., 2021; Anwar et al., 2022; Kulkarni et al., 2024; Shen et al., 2025). Alternative approaches include using defensive deception strategies to increase the cost of attacks through modeling uncertainty via hypergames (Jia et al., 2024) and maximizing strategic surprise via sensor revealing strategies (Udupa et al., 2024), or exploring two-way misinformation in deception games (Cho et al., 2019) and defensive deception against multiple attackers (Wan et al., 2023).

Social Simulations Social simulations involve multi-agent-based modeling of strategic interactions within conceptualized social environments. The framework proposed by (Trencsenyi et al., 2025a) employs a hypergame-based **Theory of Mind** reasoning framework. This approach has been utilized by (Trencsenyi et al., 2025b) to explore the ability of agentic large language models (LLMs) to engage in human-like reasoning during guessing games. Furthermore, (Kulkarni et al., 2025) leverages hypergame theory to model **Coalition Detection** in natural language reasoning tasks performed by LLMs.

Robotics In the realm of robotic **Control Theory**, hypergame applications, for example, focus on modeling obstructed perceptions in traffic scenarios using subjective games (Kahn et al., 2022). In contrast, (Kulkarni and Fu, 2019) aim to develop hypergame-based strategies in asymmetrical information environments between robots and their surroundings. Other studies in robotics explore the implementation of hypergames in the context of **Autonomous Vehicles**, particularly in the design of deceptive defense strategies for unmanned ground vehicles that operate under asymmetric perception conditions (Lv et al., 2024). Others address the cognitive alignment of unmanned aerial vehicles during defensive (He et al., 2022; Hai et al., 2023) or exploratory (Dharmadhikari et al., 2021) missions.

Multi-agent Systems We categorize applications related to general issues within MAS under this domain. (Aitchison et al., 2021) defines a set of hidden-role games targeting deceptive objectives and equipping agents with second-order **Theory of Mind** conceptualized as second-order hypergames. This framework allows agents to act based on misaligned perceptions of their roles and objectives, facilitating the emergence of asymmetric beliefs and belief manipulation within partially observable, mixed cooperative-competitive settings.

Communications Within the communications sector, we examine decentralized multi-user wireless networks composed of transmitters and receivers whose interactions are limited by bandwidth, channel quality, and computational resources. **Semantic communication** described a paradigm where only the most task-relevant information is transmitted while receivers utilize reasoning and reconstruction techniques to infer any missing data. In this context, (Thomas and Saad, 2024) applies hypergame theory to model and manage perceptual discrepancies among agents explicitly.

3.4.2 Hypergame Integration Methodology

During the paper selection process, we have already selected a subset of papers that satisfy the agent compatibility criteria. We then further classified these works based on the practicality of their proposed implementation, as follows.

- **Theoretical** works that extend or use hypergame theory in an application domain relying on agency. While these works do not entail a runnable system or dynamic experiments, an applicable algorithmic description and/or a practical case is provided.
- **Experimental** works entail hypergame-based applications that implement an experimental setting and use empirical evidence to prove or evaluate the effects of hypergame-based approaches.
- **Practical** applications encapsulate works that implement hypergame frameworks that are deployable and are tested in realistic simulations and real-life inspired settings.

Figure 6 reveals that even in our surveyed set of works that we filtered based on agent-compatibility theoretical papers still contribute to half of all relevant publications, while experimental and practical works share a count of 11 each. Even across

the more practical domains – cybersecurity, robotics, social simulations – theoretical papers prevail. This significant focus on the theoretical application of hypergames is well-aligned with the original design of hypergame theory as an analytical framework and with the challenges we articulated previously.

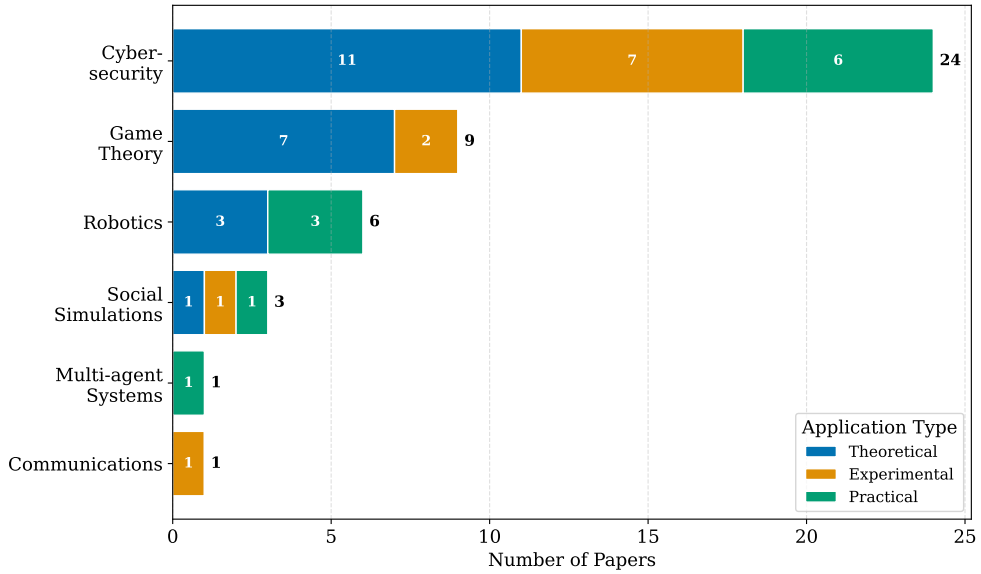

Fig. 6: Domain-specific distribution of agent-compatible hypergame papers by application type. **Theoretical** works are conceptually grounded and contain algorithmic descriptions or formalized case examples, but lack implementation or simulation. **Experimental** works implement hypergame-based models in controlled settings to evaluate their effects empirically. **Practical** applications deploy fully integrated hypergame frameworks in realistic simulations or real-world-inspired scenarios.

3.4.3 Hypergame Integration Fidelity

Next, we distinguish works based on the extent of their hypergame integration as follows:

- **Conceptual** applications refer to hypergames and are inspired by hypergame-theoretic concepts but do not incorporate an explicit hypergame model.
- **Partial** applications are inspired by hypergame-theoretic concepts but only formally integrate components – such as solution concepts or subjective perceptions – instead of instantiating the entire framework;
- **Complete** applications entail a full hypergame model integration.

As depicted by Figure 7, we can observe that 35 of surveyed papers entail a complete integration of hypergame-theoretic models, with 5 works relying on partial and

3 on conceptual implementations. Furthermore, the "complete" category is dominant across individual domains, suggesting that while the operationalization of the complex hypergame analytical frameworks is non-trivial, agent-based application benefit from high-fidelity instantiations of hypergame-theoretic models.

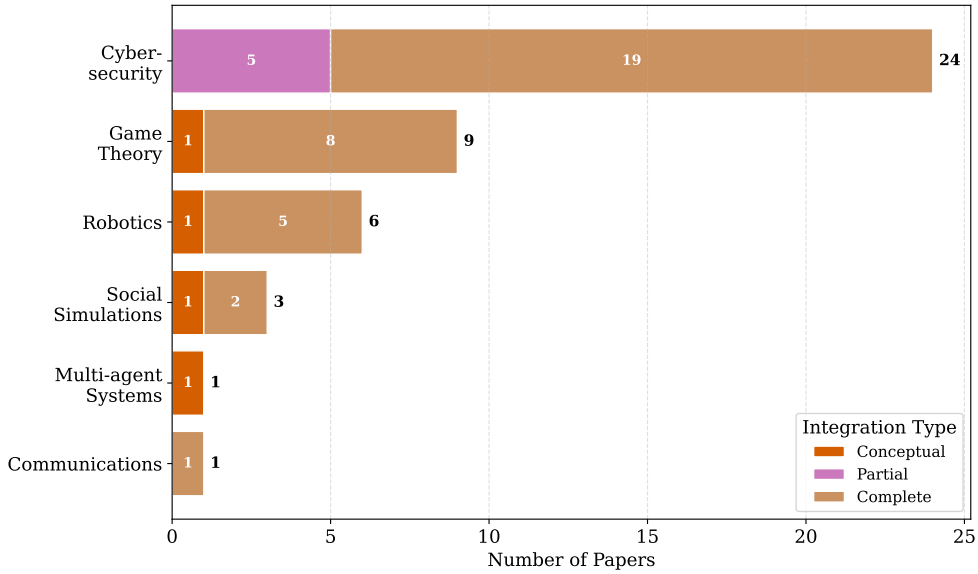

Fig. 7: Domain-specific distribution of agent-compatible hypergame papers based on the extent of model integration. **Conceptual** applications reference hypergame-theoretic ideas but do not instantiate explicit models. **Partial** applications formally incorporate some hypergame components such as subjective perceptions or equilibrium concepts, without realizing a complete framework. **Complete** applications implement fully specified hypergame models within the agentic context.

3.4.4 Hypergame Task

Finally, we classify hypergame models based on their role in the framework; what is the task the approach is seeking to solve via hypergame theory? The classification entails planning, uncertainty measure, learning, and reasoning. Models addressing multiple tasks were counted for each class. The overall distribution – as shown in Figure 8 – depicts that the majority of works utilize hypergame-theoretic models for reasoning – or decision-making in general – or complementing learning algorithms.

Table 3: Summary of agent-compatible hypergame paper classification based on model taxonomy and model application details.

	Hypergame Model Taxonomy			Hypergame Application Details				
Paper	Dynamics	Concept	Formalization	Domain[1]	Sub-domain[2]	Method	Fidelity	Task
(Jia et al., 2024)	Repeated	HNF	HEU-based	CS	DD	Practical	Partial	Reasoning, Uncertainty Measure
(Gutierrez et al., 2018)	One-shot	MLH	L-th Order	CS	ADM	Theoretical	Complete	Reasoning
(Xi and Kamhoua, 2020)	Repeated	MLH	L-th Order	CS	DD	Experimental	Complete	Reasoning
(Tang et al., 2024)	Repeated	MLH	Stackelberg	CS	ADM	Experimental	Complete	Learning
(Kulkarni and Fu, 2021b)	Repeated	MLH	Graph-based	CS	ADM	Theoretical	Complete	Reasoning
(Trencsenyi et al., 2025a)	One-shot	MLH	Flattened L-th Order	SS	TOM	Practical	Complete	Reasoning
(Kulkarni et al., 2021)	Repeated	MLH	Graph-based	CS	ADM	Theoretical	Complete	Reasoning
(Kulkarni et al., 2020)	One-shot	MLH	Graph-based	CS	DD	Theoretical	Complete	Reasoning
(Kulkarni et al., 2025)	One-shot	MLH	L-th Order	SS	CD	Theoretical	Complete	Learning
(Li et al., 2023)	Repeated	MLH	Graph-based	SS	ADM	Theoretical	Complete	Planning
(Yuan et al., 2024)	Repeated	MLH	HNE-based	GT	DC	Theoretical	Conceptual	Reasoning
(Gharesifard and Cortes, 2012)	Repeated	MLH	Graph-based	GT	EOB	Theoretical	Complete	Learning
(Gharesifard and Cortés, 2010)	Repeated	MLH	Graph-based	GT	EOB	Theoretical	Complete	Learning
(Gharesifard and Cortes, 2011a)	Repeated	MLH	Graph-based	GT	EOB	Theoretical	Complete	Learning, Uncertainty Measure
(Wan et al., 2021)	Repeated	HNF	Stackelberg, HEU-based	CS	DD	Practical	Partial	Reasoning, Uncertainty Measure
(Anwar et al., 2022)	Repeated	HNF	HEU-based	CS	DD	Practical	Partial	Reasoning
(Li et al., 2024)	Repeated	MLH	Flattened, L-th Order	GT	P-E	Experimental	Complete	Reasoning, Learning
(House and Cybenko, 2010)	Repeated	HNF	HNF	CS	ADM	Experimental	Complete	Learning
(Thomas and Saad, 2024)	Repeated	MLH	Stackelberg	COM	SC	Experimental	Complete	Learning
(Dharmadhikari et al., 2021)	Repeated	MLH	Perceptual	ROB	AV	Practical	Conceptual	Planning
(Kahn et al., 2022)	One-shot	MLH	L-th Order	ROB	CT	Practical	Complete	Uncertainty Measure
(Kulkarni et al., 2024)	Repeated	MLH	Graph-based	CS	DD	Experimental	Complete	Reasoning
(Shen et al., 2025)	Repeated	MLH, HNF	L-th Order, HEU-based	CS	DD	Practical	Complete	Reasoning
(Thukkaraju et al., 2023)	Repeated	HNF	HEU-based	CS	ADM	Experimental	Partial	Reasoning, Learning
(He et al., 2022)	One-shot	MLH	Stackelberg	ROB	AV	Theoretical	Complete	Reasoning

[1]CS: Cybersecurity, SS: Social Simulations, GT: Game Theory, COM: Communications, ROB: Robotics

[2]DD: Defensive Deception, ADM: Attack-Defense Modeling, TOM: Thoery of Mind, CD: Coalition Detection, EOB: Evolution of Beliefs, P-E: Pursuit-Evasion, SC: Semantic Communication, AV: Autonomous Vehicles, CT: Control Theory

Table 4: (Cont.) Summary of agent-compatible hypergame paper classification based on model taxonomy and model application details.

	Hypergame Model Taxonomy			Hypergame Application Details				
Paper	Dynamics	Concept	Formalization	Domain[1]	Sub-domain[2]	Method	Fidelity	Task
(Bakker et al., 2020)	Repeated	MLH	L-th Order, HNE-based	CS	ADM	Theoretical	Complete	Reasoning
(Gharesifard and Cortés, 2011b)	Repeated	MLH	Graph-based	GT	EOB	Theoretical	Complete	Reasoning, Learning
(Aitchison et al., 2021)	Repeated	MLH	Bayesian L-th Order	MAS	TOM	Practical	Conceptual	Reasoning, Learning
(Bakker et al., 2021)	One-shot	MLH	L-th Order	CS	ADM	Theoretical	Complete	Reasoning
(Cho et al., 2019)	Repeated	HNF	HEU-based	CS	DD	Experimental	Complete	Reasoning, Uncertainty Measure
(Kulkarni and Fu, 2019)	One-shot	MLH	L-th Order	ROB	CT	Theoretical	Complete	Reasoning, Planning
(Lv et al., 2024)	Repeated	MLH	Graph-based	ROB	AV	Practical	Complete	Reasoning
(Shi et al., 2023)	Repeated	MLH	L-th Order	GT	SG	Experimental	Complete	Planning
(Udupa et al., 2024)	Repeated	MLH	Graph-based	CS	DD	Theoretical	Complete	Reasoning
(Wan et al., 2023)	Repeated	MLH, HNF	L-th Order, HEU-based	CS	DD	Practical	Complete	Reasoning, Learning
(Cheng et al., 2022)	One-shot	MLH	Stackelberg, HNE-based	CS	ADM	Theoretical	Complete	Reasoning
(Gharesifard and Cortés, 2014)	Repeated	MLH	Graph-based	CS	ADM	Theoretical	Complete	Reasoning, Learning
(Gharesifard and Cortés, 2011c)	Repeated	MLH	Graph-based	GT	EOB	Theoretical	Complete	Reasoning, Learning
(Yoon et al., 2025)	One-shot	HNF	HEU-based	CS	ADM	Practical	Partial	Reasoning, Uncertainty Measure
(Kulkarni and Fu, 2021a)	One-shot	MLH	Graph-based	CS	ADM	Theoretical	Complete	Reasoning
(Trencsenyi et al., 2025b)	One-shot	MLH	Flattened L-th Order	SS	TOM	Experimental	Conceptual	Reasoning
(Hai et al., 2023)	One-shot	MLH	Stackelberg	ROB	AV	Theoretical	Complete	Reasoning
(Das et al., 2020)	Repeated	MLH	L-th Order	CS	TD	Experimental	Complete	Reasoning
(Song et al., 2009)	One-shot	MLH	L-th Order	GT	FUZ	Theoretical	Complete	Reasoning

[1]CS: Cybersecurity, GT: Game Theory, MAS: Multi-agent Systems, ROB: Robotics, SS: Social Simulations

[2]ADM: Attack-Defense Modeling, EOB: Evolution of Beliefs, TOM: Thoery of Mind, DD: Defensive Deception, CT: Control theory, AV: Autonomous Vehicles, TD: Trojan Detection, FUZ: Fuzzy Game Theory

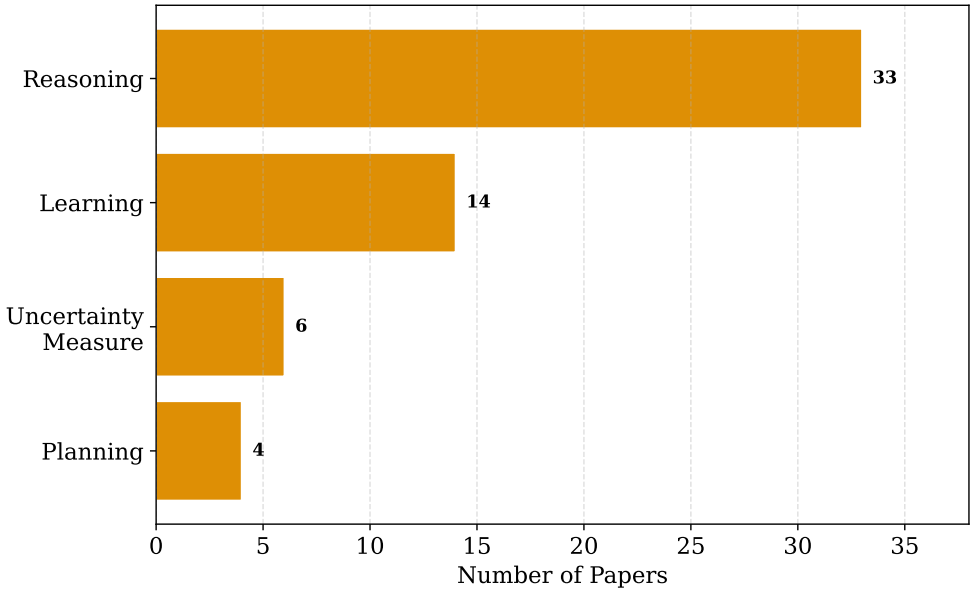

Fig. 8: Distribution of surveyed papers by primary computational task implemented by hypergame-theoretic models.

4 Discussion of Survey Results

In the preceding sections, we established a comprehensive taxonomy of hypergame models and explored various applications that utilize these models. The full classification of the 44 reviewed papers is documented in Tables 3 and 4. In this section, we reflect on emerging patterns and structural gaps identified through our systematic review of hypergame-theoretic literature and the classifications disclosed in the previous sections, aiming to contextualize the current state of the field and identify fruitful directions for future research.

4.1 Hypergame Usage Patterns

In this section, we analyze emerging patterns in the design and usage of hypergame models across agentic applications. Drawing on the classification dimensions outlined earlier, we identify recurring choices in model structures, domain integration, and computational tasks. These patterns offer practical insights for researchers aiming to deploy hypergame-theoretic mechanisms in dynamic multi-agent systems.

4.1.1 Overview of Model Usage

An evaluation of the core concept (multi-level hypergame vs. HNF) usage across surveyed works at a domain-level decomposition (Figure 9) reveals that multi-level hypergames are the dominant framework across all categories, and only cybersecurity works entail implementations of HNF or hybrid models.

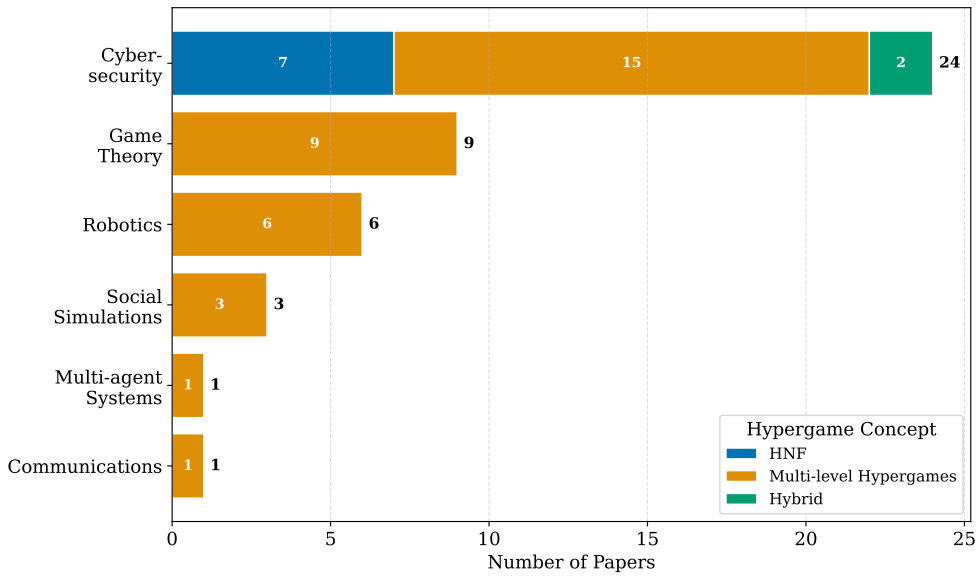

Fig. 9: Distribution of core hypergame concepts implemented per application domain. **HNF** entails works inspired by Vane's hypergame normal form (Vane and Lehner, 2000), **Multi-level Hypergames** are based on Bennet's original concept (Bennett, 1980) or Wang et al.'s extended formalism (Wang et al., 1988). **Hybrid** approaches use concepts from both main concepts.

At a lower-level view, graph-based hypergames and L-th order hierarchical hypergames dominate (Figure 4). Graph-based models in particular stand out as the most popular implementation, often supporting spatial planning and adversarial reasoning, especially in cybersecurity and general game-theoretic applications (Figure 10).

However, when theoretical applications are omitted, shifting the focus toward experimental and practical implementations that are closer to deployable multi-agent systems, the distribution changes considerably. As depicted in Figure 11, 8 out of the 14 works we classified as practical entail simplified hypergame-theoretic concepts – perceptual games, flattened L-th order hypergames or HEU – instead of integrating and populating models entirely according to the complete hypergame analyses frameworks. These findings suggest, that while practical, real-world applications can benefit from hypergame-theoretic models, their integration must entail concretizations of the primarily analytical frameworks. Figures 6 and 12 paint a similar picture, where practical application from domains more aligned with real-world applicability – cybersecurity, robotics, social simulations and general MAS – works seem to benefit most from the simplified models.

Meanwhile, HNF has seen limited use in agentic contexts. Only 9 works adopt HNF-based solutions (Figure 9), with all situated in the cybersecurity domain. This suggests

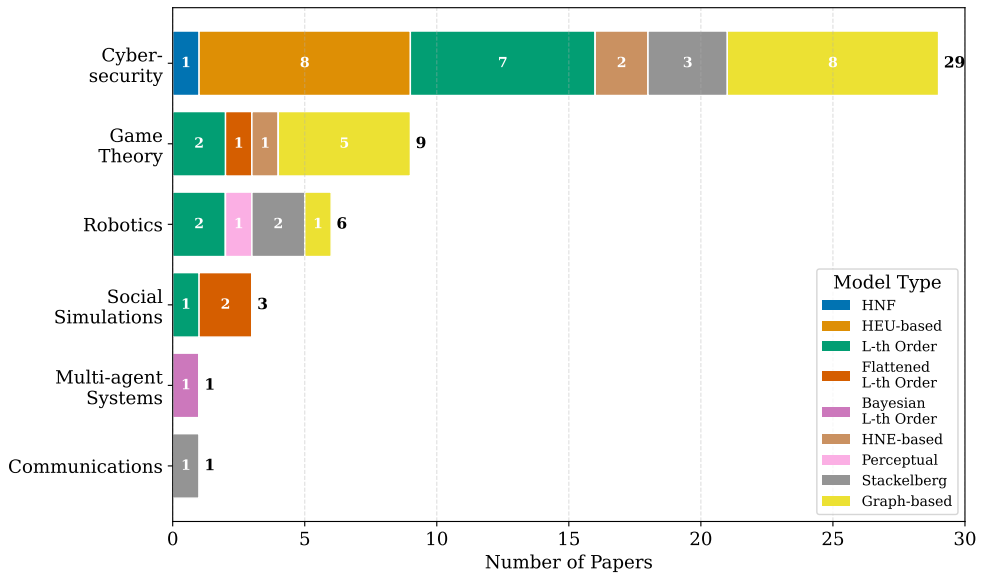

Fig. 10: Distribution of hypergame model types across application domains. Hybrid models that span multiple categories are counted in each relevant type.

that HNF's fixed-matrix structure may be too rigid or low-level for modern agent-based modeling, where higher-order and domain-specific reasoning demands greater representational flexibility.

4.1.2 Task-Model Alignment

As we can see on Figure 13, reasoning is the most frequent application, with 33 papers using hypergames to support decision-making under misaligned perceptions. Hypergame-based learning tasks are present in about a third of the surveyed papers, most frequently associated with graph-based and HEU-based models, which offer scalable formulations that lend themselves to integration with learning algorithms. Apart from one paper, estimating uncertainty appears as a secondary task, complementing reasoning or learning, with mostly probabilistic, HEU-based approaches. The limited presence of hypergames in high-level planning tasks may reflect the absence of expressive, agent-compatible formalisms akin to epistemic planning languages.

To complement these findings, we evaluate how different task-based works are composed of theoretical, experimental and practical model applications (Figure 14) and revisited the distribution of hypergame model types per tasks (Figures 6, 12). While practical applications are the majority for deriving hypergame-based uncertainty measures, for planning, learning and reasoning-related tasks purely theoretical and more applicable works have an equal split.

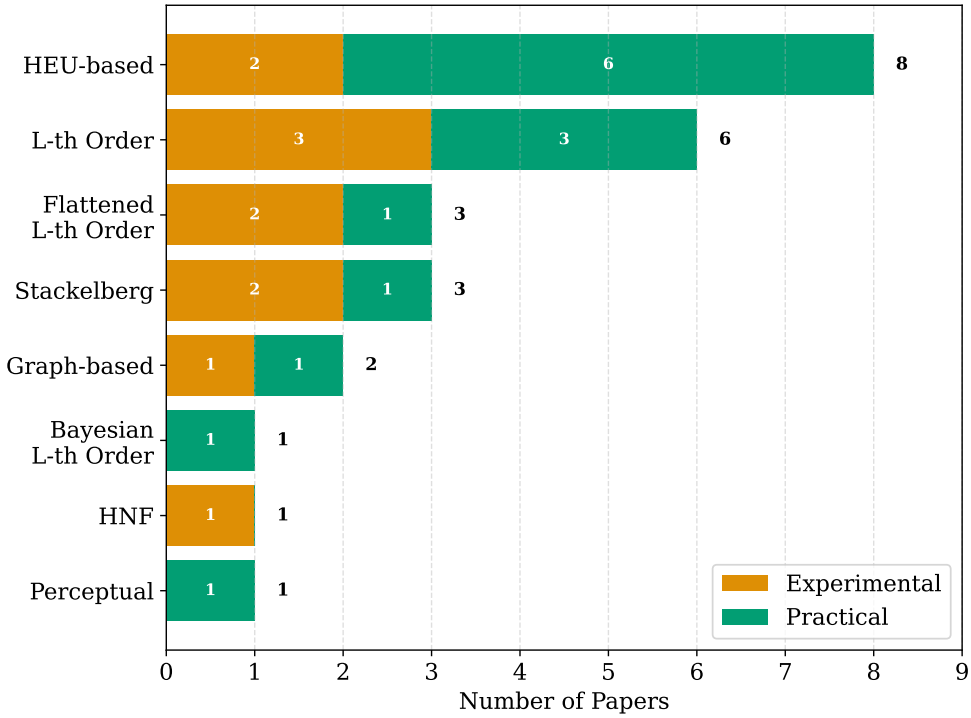

Fig. 11: Distribution of experimental and practical hypergame applications across model types.

4.1.3 Domain-Model Associations

When decomposed by application domain (Figure 10), distinct preferences in model types emerge:

- Cybersecurity leads in diversity and volume, encompassing both L-th order and HNF-based models, as well as graph-based and Stackelberg hypergames or hypergame equilibrium-based appraoches. The prevalence of deception-centric scenarios appears to benefit from nested and multi-layered perception modeling, making hypergames a natural fit.
- General game-theoretic works entail mostly theoretical but high-fidelity, complete integrations of hypergame models.
- Social simulations and multi-agent system studies predominantly use flattened L-th order and L-th order hypergames, highlighting a focus on recursive reasoning rather than spatial deception. This aligns with their frequent overlap with theory of mind literature and epistemic modeling.
- Robotics and communications rely primarily on subjective perceptual representations, which can encode environmental uncertainties and cognitive bias, particularly in navigation and defense tasks.

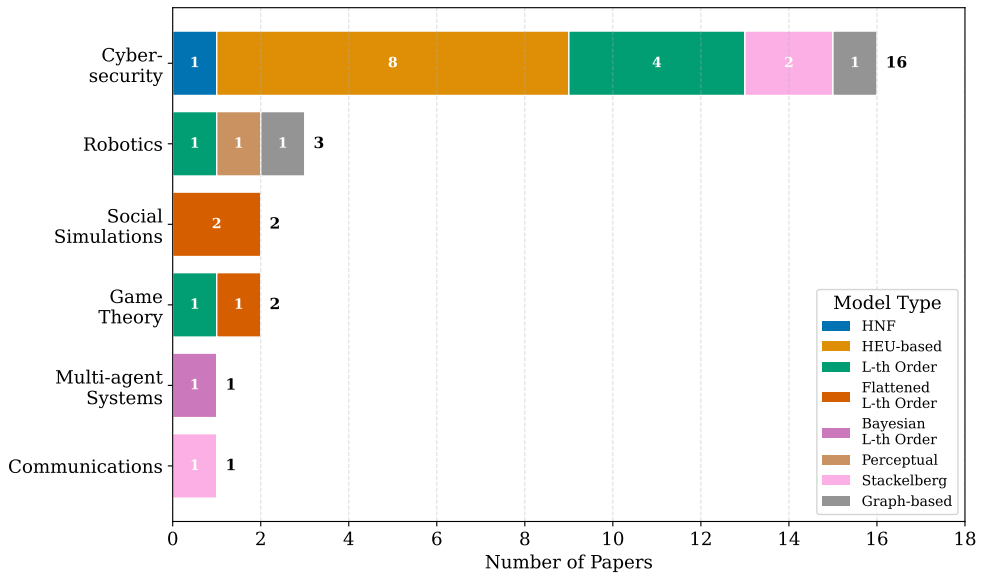

Fig. 12: Distribution of hypergame model across application domains, omitting theoretical applications.

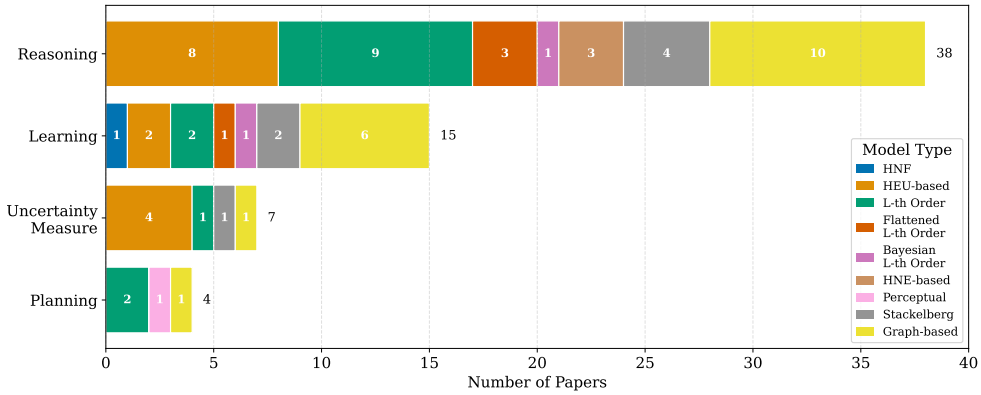

Fig. 13: Distribution of model types per hypergame task.

4.2 Agent-based Hypergame Languages

While the extended formalisms by (Wang et al., 1988) and (Vane and Lehner, 2000) concretized hypergame frameworks and provided guidelines for hypergame analysis, there is no unified formal language for reasoning about hypergames. Motivated by this gap, hypergame markup language (HML) was developed by (Brumley, 2003) to formally encode hypergame models for his hypergame analysis software: HYPANT. HML supports subjective subgames based on player perception but is limited to static

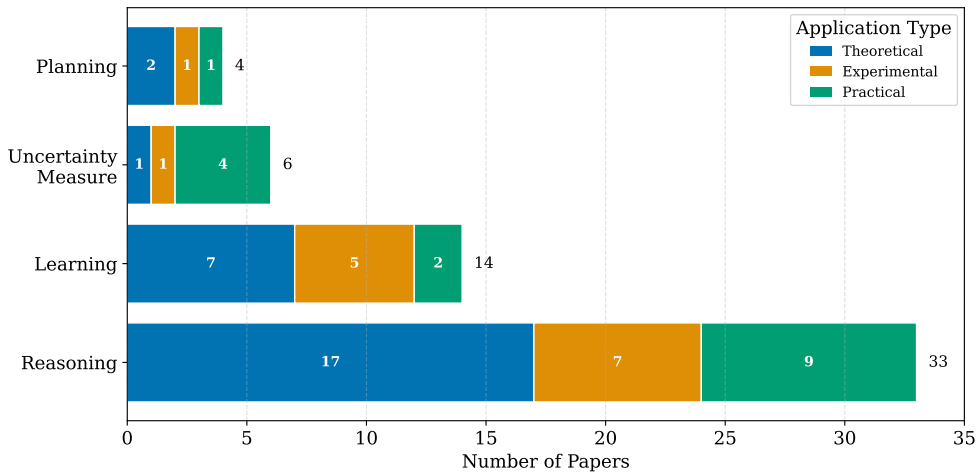

Fig. 14: Distribution of hypergame task with composition of theoretical, experimental and practical applications.

utilities and does not support functional utility definitions or dynamic belief updates. Another hypergame analysis software, HAT (Gibson, 2013), relies on XML-based structured representation of hypergames. While HAT supports iterative runs and belief updates, it relies on populating Vane's hypergame normal form representation-based structures. (Kovach et al., 2015) also mention the Security Policy Assistant (SPA) developed by Vane to assist in sensitive decision-making, such as evaluating foreign disclosure of classified information. SPA is not publicly available and we could not find any material for a more detailed analysis.

Despite the aforementioned tools and preliminary modeling languages, an agent-based hypergame modeling language or simulation platform has not emerged. While we encountered various works in our survey that integrate hypergame-theoretic models with linear-temporal-logic-based planning, they do not define a hypergame-specific description language. Future hypergame modeling efforts may take inspiration from alternative frameworks that address information asymmetry and nested beliefs to develop a hypergame-theoretic formal language grounded in logic, such as epistemic logic model that enable dynamic belief updates and planning with iterated knowledge (Baltag and Smets, 2008; Bolander and Andersen, 2011), GDL-III providing a formal specification language for general game playing with games of incomplete information (Thielscher, 2017), or TOMA extending epistemic logic to formalize the theory of mind (Erdogan et al., 2025).

5 Related Work

While Bennett initially developed hypergame theory as an analytical tool for structuring complex strategic scenarios, subsequent research has explored more dynamic

applications. (Bennett et al., 1981) demonstrated the use of hypergame theory in real-time decision-making during a shipping crisis. Later, the development of Vane's HNF provided a more methodological version with potential applications in dynamic, agent-based models. The potential is articulated in his paper addressing decision-theoretic agents (Vane, 2000) and his brief, high-level review of hypergames (Vane, 2006), where the connection to the Observe-Orient-Decide-Act paradigm was also established – a concept well-aligned with human-cognition-inspired agent theories such as the belief-desire-intention (BDI) architecture (Tweedale et al., 2007; Simari and Parsons, 2011). Jointly, including the additional formalism provided by (Wang et al., 1988), while these advancements of hypergame theory offer a richly formal and generalizable framework, up to this point, this line of research remains unconnected to the agent-based and multi-agent systems literature.

(Kovach et al., 2015) presented a broad review of hypergame theory as a framework for modeling conflict, misperception, and deception, particularly in adversarial domains like military planning and cyber defense. Their review emphasizes the theoretical basis of hypergames and highlights how layered beliefs and misaligned perceptions give rise to strategic opportunities unavailable in classical game theory or decision theory. However, their review primarily focused on illustrative, mostly analytical examples rather than methodological analyses.

(Zhu et al., 2021) offered a dual-framework perspective, reviewing defensive deception from both machine learning and game-theoretic viewpoints. Although hypergame theory is featured in their survey through selected analytical and practical examples, the coverage remains limited and application-specific, without substantial linkage to broader multi-agent systems research. Complementing this work, (Khalid et al., 2023) presented a systematic review of game-theoretic defense models against advanced persistent threats. While not dedicated to hypergames and, as such, lacking formal descriptions of hypergame models, their review acknowledges the value of hypergame theory in modeling adversarial perturbations and covers three additional hypergame-based works on defensive deception. Crook's doctoral dissertation (Crook, 2024) represents a unique contribution bridging computability theory and hypergames, involving an exploration of how attack graphs in cybersecurity can be recast as hypergames. While preliminary, this work proposes several directions for formalizing adaptive, layered-belief hypergames in cybersecurity with links to computable analysis. Finally, (Fugate and Ferguson-Walter, 2019) addressed hypergame theory in their review of game-theoretic models in cyber deception. While their discussion is brief and only on a high level, the paper explicitly recognizes that attacker–defender interactions often exhibit asymmetries traditional games cannot capture. The consideration of hypergames for the adaptability of defender strategies progresses the practical generalizability of hypergame theory.

To our knowledge, no prior work provides a systematic survey of hypergame theory from the perspective of multi-agent systems. Existing reviews typically focus on early decision-theoretic formalisms, defensive deception in cyber-physical systems, or specific ad-hoc and post-hoc case studies. Our survey addresses this gap by:

- Providing a rigorous classification framework explicitly tailored to multi-agent systems, emphasizing works that implement and apply hypergame theory within agent-compatible contexts;
- Presenting a structured taxonomy of hypergame models across multiple dimensions, including formal hypergame structures, agent architectures, domain-specific implementations, and computational tasks;
- Identifying key patterns in hypergame integration, such as the prevalence of cybersecurity-related hypergame-theoretic applications, while also highlighting and discussing reasons for the limited use of frameworks like Hypergame Normal Form (HNF) in other domains;
- Mapping the explicit relationship between hypergame model types and their deployment in reasoning, planning, learning, and uncertainty management tasks, thereby underscoring the alignment between hypergame formalisms and their intended agentic functionalities;
- Demonstrating the evolving landscape of hypergame theory, shifting from historically concentrated theoretical analyses toward more diverse, practical deployments across multiple MAS-relevant domains;
- Identifying structural gaps and open research challenges, particularly in hypergame-based epistemic modeling, human–AI interaction, and high-level planning frameworks.

By synthesizing these insights, our survey not only bridges existing knowledge gaps but also provides a clear roadmap for future hypergame-theoretic research, particularly within dynamic, agent-driven multi-agent systems. The outlined trends and classifications serve as actionable references, encouraging researchers to leverage hypergames in practical MAS applications and inspiring further methodological innovation within the hypergame theory community.

6 Conclusions

In this work, we conducted a systematic review of hypergame theory, focusing on its adaptation and application within multi-agent systems (MAS). We surveyed 44 papers spanning multiple domains and analyzed how hypergame models have been integrated into agent-based frameworks, examining trends, formalizations, and application patterns. Our review highlighted both the theoretical strengths of hypergame theory – particularly its ability to represent nested beliefs, misaligned perceptions, and subjective representations of strategic interactions – and the practical challenges of deploying such models in dynamic MAS environments.

Despite early efforts such as HML (Brumley, 2003) and HAT (Gibson, 2013), there remains no formal language or simulation framework explicitly designed for modeling agent-based hypergames. Given the identified lack of a unified representational formalism and the need for structured support for recursive belief modeling and dynamic perception updates, future work should prioritize the development of dedicated hypergame description languages tailored to agent-based applications. Inspiration may be drawn from planning frameworks rooted in epistemic logic (Baltag and Smets, 2008; Bolander and Andersen, 2011), general game-playing languages like

GDL-III (Thielscher, 2017), or formal ToM modeling tools such as TOMA (Erdogan et al., 2025).

Across the surveyed domains, hypergames have been successfully employed to model nested beliefs, misaligned perceptions, and deception; however, few works explicitly link these models to established cognitive agent architectures. As recursive reasoning becomes increasingly central in agent design, integrating hypergame theory with human-inspired frameworks such as BDI (Tweedale et al., 2007; Simari and Parsons, 2011) could enrich symbolic reasoning and offer interpretable mechanisms for belief formation and adaptation in dynamic environments. Furthermore, while we encountered only a handful of hypergame-based approaches addressing large language models (LLMs), the growing prevalence of human–agent (Strachan et al., 2024) and agent–agent misalignment (Kierans et al., 2025), along with the application of theory of mind in agentic systems (Bosse et al., 2011; Rocha et al., 2023; Erdogan et al., 2025), suggests promising opportunities for hypergame theory to formalize ToM-like models capable of capturing and mitigating agent misalignment.

Finally, several surveyed works highlighted hybrid reasoning approaches, such as Bayesian L-th order hypergames, yet few explored the design space that combines subjective perception modeling with probabilistic inference. Current HNF-based approaches have not yet provided a general solution for dynamically populating the HNF matrix. Future research could explore hybrid agent architectures that integrate hypergame structures for belief representation with probabilistic reasoning modules for decision-making, enabling agents to better navigate uncertain, misperceived, or adversarial contexts.

We hope that this survey will guide and inspire future researchers seeking to model misperceptions, nested beliefs, and recursive reasoning in multi-agent-based applications. By identifying trends, gaps, and trajectories in the literature, we aim to catalyze the development of hypergame-based AI models that are both theoretically grounded and practically applicable.

7 Declarations

Conflict of interest. The authors declare no conflict of interest.

Funding. This work was supported by a Leverhulme Trust International Professorship Grant (LIP-2022-001).

Acknowledgement. The authors would like to express their gratitude to David Levine for his constructive comments. The first author would also like to thank him for his supervision and the fruitful discussion of game-theoretic and hypergame-theoretic materials.

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a

credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aitchison, M., Benke, L., Sweetser, P.: Learning to deceive in multi-agent hidden role games. In: Sarkadi, S., Wright, B., Masters, P., McBurney, P. (eds.) Deceptive AI, pp. 55–75. Springer, Cham (2021)

Anwar, A.H., Zhu, M., Wan, Z., Cho, J.-H., Kamhoua, C.A., Singh, M.P.: Honeypot-based cyber deception against malicious reconnaissance via hypergame theory. In: GLOBECOM 2022 - 2022 IEEE Global Communications Conference, pp. 3393–3398 (2022). https://doi.org/10.1109/GLOBECOM48099.2022.10000813

Bolander, T., Andersen, M.B.: Epistemic planning for single-and multi-agent systems. Journal of Applied Non-Classical Logics **21**(1), 9–34 (2011) https://doi.org/10.3166/jancl.21.9-34

Bakker, C., Bhattacharya, A., Chatterjee, S., Vrabie, D.L.: Learning and information manipulation: Repeated hypergames for cyber-physical security. IEEE Control Systems Letters **4**(2), 295–300 (2020) https://doi.org/10.1109/LCSYS.2019.2925681

Bakker, C., Bhattacharya, A., Chatterjee, S., Vrabie, D.L.: Metagames and hypergames for deception-robust control. ACM Trans. Cyber-Phys. Syst. **5**(3) (2021) https://doi.org/10.1145/3439430

Bennett, P.G., Dando, M.R.: Complex strategic analysis: A hypergame study of the fall of france. The Journal of the Operational Research Society **30**(1), 23–32 (1979)

Bennett, P.: Toward a theory of hypergames. Omega **5**(6), 749–751 (1977) https://doi.org/10.1016/0305-0483(77)90056-1

Bennett, P.G.: Hypergames: Developing a model of conflict. Futures **12**(6), 489–507 (1980) https://doi.org/10.1016/0016-3287(80)90005-1

Bennett, P., Huxham, C., Dando, M.: Shipping in crisis: A trial run for 'live' application of the hypergame approach. Omega **9**(6), 579–594 (1981) https://doi.org/10.1016/0305-0483(81)90046-3

Batzilis, D., Jaffe, S., Levitt, S., List, J.A., Picel, J.: Behavior in strategic settings: Evidence from a million rock-paper-scissors games. Games **10**(2) (2019) https://doi.org/10.3390/g10020018

Bosse, T., Memon, Z.A., Treur, J.: A recursive bdi agent model for theory of mind and its applications. Applied Artificial Intelligence **25**(1), 1–44 (2011) https://doi.

org/10.1080/08839514.2010.529259

Brumley, L.: Hypant: a hypergame analysis tool. PhD thesis, Monash University (2003)

Baltag, A., Smets, S.: A qualitative theory of dynamic interactive belief revision. Logic and the foundations of game and decision theory (LOFT 7) **3**, 9–58 (2008) https://doi.org/10.1007/978-3-319-20451-2_39

Cheng, Z., Chen, G., Hong, Y.: Single-leader-multiple-followers stackelberg security game with hypergame framework. IEEE Transactions on Information Forensics and Security **17**, 954–969 (2022) https://doi.org/10.1109/TIFS.2022.3155294

Camerer, C.F., Ho, T.-H., Chong, J.-K.: A Cognitive Hierarchy Model of Games*. The Quarterly Journal of Economics **119**(3), 861–898 (2004) https://doi.org/10.1162/0033553041502225

Crook, T.: Computable analysis and game theory: From foundations to applications. Phd thesis, Swansea University, Wales, UK (2024). https://doi.org/10.23889/SUThesis.67150 . https://doi.org/10.23889/SUThesis.67150

Cho, J.-H., Zhu, M., Singh, M.: In: Al-Shaer, E., Wei, J., Hamlen, K.W., Wang, C. (eds.) Modeling and Analysis of Deception Games Based on Hypergame Theory, pp. 49–74. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-02110-8_4

Dharmadhikari, M., Deshpande, H., Dang, T., Alexis, K.: Hypergame-based adaptive behavior path planning for combined exploration and visual search. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 269–275 (2021). https://doi.org/10.1109/ICRA48506.2021.9561451

Das, T., Eldosouky, A.R., Sengupta, S.: Think smart, play dumb: Analyzing deception in hardware trojan detection using game theory. In: 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pp. 1–8 (2020). https://doi.org/10.1109/CyberSecurity49315.2020.9138897

Dekel, E., Siniscalchi, M.: Chapter 12 - epistemic game theory. In: Young, H.P., Zamir, S. (eds.) Handbook of Game Theory with Economic Applications vol. 4, pp. 619–702. Elsevier, North Holland (2015). https://doi.org/10.1016/B978-0-444-53766-9.00012-4

Erdogan, E., Dignum, F., Verbrugge, R., Yolum, P.: Toma: Computational theory of mind with abstractions for hybrid intelligence. Journal of Artificial Intelligence Research **82**, 285–311 (2025) https://doi.org/10.1613/jair.1.16402

Fugate, S., Ferguson-Walter, K.: Artificial intelligence and game theory models for defending critical networks with cyber deception. AI Magazine **40**(1), 49–62 (2019) https://doi.org/10.1609/aimag.v40i1.2849

Gutierrez, C.N., Almeshekah, M.H., Bagchi, S., Spafford, E.H.: A hypergame analysis for ersatzpasswords. In: Janczewski, L.J., Kutyłowski, M. (eds.) ICT Systems Security and Privacy Protection, pp. 47–61. Springer, Cham (2018)

Gharesifard, B., Cortés, J.: Evolution of the perception about the opponent in hypergames. In: 49th IEEE Conference on Decision and Control (CDC), pp. 1076–1081 (2010). https://doi.org/10.1109/CDC.2010.5718174

Gharesifard, B., Cortes, J.: Exploration of misperceptions in hypergames. In: 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1565–1570 (2011). https://doi.org/10.1109/Allerton.2011.6120354

Gharesifard, B., Cortés, J.: Learning of equilibria and misperceptions in hypergames with perfect observations. In: Proceedings of the 2011 American Control Conference, pp. 4045–4050 (2011). https://doi.org/10.1109/ACC.2011.5991206

Gharesifard, B., Cortés, J.: Stealthy strategies for deception in hypergames with asymmetric information. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference, pp. 5762–5767 (2011). https://doi.org/10.1109/CDC.2011.6160979

Gharesifard, B., Cortes, J.: Evolution of players' misperceptions in hypergames under perfect observations. IEEE Transactions on Automatic Control **57**(7), 1627–1640 (2012) https://doi.org/10.1109/TAC.2011.2180113

Gharesifard, B., Cortés, J.: Stealthy deception in hypergames under informational asymmetry. IEEE Transactions on Systems, Man, and Cybernetics: Systems **44**(6), 785–795 (2014) https://doi.org/10.1109/TSMC.2013.2277695

Gibbons, R.: A Primer in Game Theory. Pearson Education Limited, Harlow, Essex, United Kingdom (1992)

Gibson, A.S.: Applied hypergame theory for network defense. M.s. thesis, Air Force Institute of Technology (2013). Theses and Dissertations

Hansson, S.O.: Decision theory: An overview. In: International Encyclopedia of Statistical Science (2011)

Harsanyi, J.: Games with incomplete information played by 'bayesian' players, parts i-iii. Management Science **14**, 159–182320334486502 (1967-1968)

House, J.T., Cybenko, G.: Hypergame theory applied to cyber attack and defense. In: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense IX, vol. 7666, pp. 39–49 (2010). https://doi.org/10.1117/12.852338 . SPIE

He, B., Gao, J., Guan, X., Cheng, Z.: Leader-follower security games in uav communication with deception. In: Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), vol. 12506, pp. 347–352 (2022). https://doi.org/10.1117/12.2662028 . SPIE

Hai, X., Ma, X., Zhang, W.: The influence of the uncertainty on drones' cognition in unmanned aerial vehicles communication network. In: Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence. ICCAI '23, pp. 589–594. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3594315.3594376

Jia, G., Wang, X., Liu, H., Wu, H.: A deception defense strategy based on game theory. In: 2024 IEEE International Conference on Security, Privacy, Anonymity in Computation and Communication and Storage (SpaCCS), pp. 27–33 (2024). https://doi.org/10.1109/SpaCCS63173.2024.00011

Khalid, M.N.A., Al-Kadhimi, A.A., Singh, M.M.: Recent developments in game-theory approaches for the detection and defense against advanced persistent threats (apts): A systematic review. Mathematics **11**(6) (2023) https://doi.org/10.3390/math11061353

Kulkarni, A.N., Cohen, M.S., Kamhoua, C.A., Fu, J.: Integrated Resource Allocation and Strategy Synthesis in Safety Games on Graphs with Deception (2024). https://arxiv.org/abs/2407.14436

Kulkarni, A.N., Fu, J.: Opportunistic synthesis in reactive games under information asymmetry. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 5323–5329 (2019). https://doi.org/10.1109/CDC40024.2019.9029851

Kulkarni, A.N., Fu, J.: Synthesis of deceptive strategies in reachability games with action misperception. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI'20 (2021)

Kulkarni, A.N., Fu, J.: A theory of hypergames on graphs for synthesizing dynamic cyber defense with deception. In: Kamhoua, C.A., Kiekintveld, C.D., Fang, F., Zhu, Q. (eds.) Game Theory and Machine Learning for Cyber Security, pp. 97–112. John Wiley & Sons, USA (2021). https://doi.org/10.1002/9781119723950.ch6

Kulkarni, A.N., Fu, J., Luo, H., Kamhoua, C.A., Leslie, N.O.: Decoy allocation games on graphs with temporal logic objectives. In: Zhu, Q., Baras, J.S., Poovendran, R., Chen, J. (eds.) Decision and Game Theory for Security, pp. 168–187. Springer, Cham (2020)

Kierans, A., Ghosh, A., Hazan, H., Dori-Hacohen, S.: Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment. Proceedings of the AAAI Conference on Artificial Intelligence **39**(26), 27365–27373 (2025) https://doi.org/10.1609/aaai.v39i26.34947

Kovach, N.S., Gibson, A.S., Lamont, G.B.: Hypergame theory: A model for conflict, misperception, and deception. Game Theory **2015**(1), 570639 (2015) https://doi.org/10.1155/2015/570639

Kulkarni, A.N., Liu, A., Gaglione, J.-R., Fried, D., Topcu, U.: Dynamic coalition structure detection in natural-language-based interactions. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. AAMAS '25, pp. 1227–1234. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2025)

Kulkarni, A.N., Luo, H., Leslie, N.O., Kamhoua, C.A., Fu, J.: Deceptive labeling: Hypergames on graphs for stealthy deception. IEEE Control Systems Letters **5**(3), 977–982 (2021) https://doi.org/10.1109/LCSYS.2020.3008078

Kahn, M., Sarkar, A., Czarnecki, K.: I know you can't see me: Dynamic occlusion-aware safety validation of strategic planners for autonomous vehicles using hypergames. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 11202–11208 (2022). https://doi.org/10.1109/ICRA46639.2022.9812041

Lv, P., Li, S., Yin, X.: Optimal deceptive strategy synthesis for autonomous systems under asymmetric information. IEEE Transactions on Intelligent Vehicles **9**(10), 6108–6121 (2024) https://doi.org/10.1109/TIV.2024.3362585

Li, L., Ma, H., Kulkarni, A.N., Fu, J.: Dynamic hypergames for synthesis of deceptive strategies with temporal logic objectives. IEEE Transactions on Automation Science and Engineering **20**(1), 334–345 (2023) https://doi.org/10.1109/TASE.2022.3150167

Luce, R.D., Raiffa, H.: Games and Decisions: Introduction and Critical Survey. Dover Books on Mathematics. Dover Publications, United Kingdom (1989)

Li, Y., Zhang, H., Liang, S.: Hyper q-learning algorithm for pursuit-evasion games with misperceptions. In: 2024 IEEE International Conference on Unmanned Systems (ICUS), pp. 1317–1322 (2024). https://doi.org/10.1109/ICUS61736.2024.10839768

Myerson, R.B.: An introduction to game theory. Discussion Papers 623, Northwestern University, Center for Mathematical Studies in Economics and Management Science (1984)

Nagel, R.: Unraveling in guessing games: An experimental study. The American Economic Review **85**(5), 1313–1326 (1995)

Nash, J.F.: Equilibrium points in n-person games. Proceedings of the National Academy of Sciences of the United States of America **36**(1), 48–49 (1950)

Osborne, M., Rubinstein, A.: A Course in Game Theory vol. 1, 1st edn. The MIT Press, Cambridge,Massachusetts (1994)

Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978) https://doi.org/10.1017/S0140525X00076512

Rasmusen, E.: Games and information. an introduction to game theory. Theory and Decision **29**(2), 161 (1990)

Rocha, M., Silva, H.H., Morales, A.S., Sarkadi, S., Panisson, A.R.: Applying theory of mind to multi-agent systems: A systematic review. In: Naldi, M.C., Bianchi, R.A.C. (eds.) Intelligent Systems, pp. 367–381. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-45368-7_24

Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., *et al.*: Testing theory of mind in large language models and humans. Nature Human Behaviour **8**(7), 1285–1295 (2024) https://doi.org/10.1038/s41562-024-01882-z

Sasaki, Y.: In: Metcalf, G.S., Kijima, K., Deguchi, H. (eds.) Multi-agent Decision System, pp. 337–352. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-0720-5_49

Shi, C., Han, S., Fu, J.: Quantitative planning with action deception in concurrent stochastic games. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. AAMAS '23, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2023)

Simon, H.A.: Theories of bounded rationality. In: McGuire, C.B., Radner, R. (eds.) Decision and Organization, pp. 161–176. North-Holland, Amsterdam (1972)

Song, Y., Li, W., Zeng, X.: λ-equilibrium of hypergame with fuzzy strategy and preference perceptions. In: 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, vol. 2, pp. 641–645 (2009). https://doi.org/10.1109/ICICISYS.2009.5358298

Simari, G.I., Parsons, S.D.: An Empirical Comparison of Models, pp. 11–25. Springer, New York, NY (2011). https://doi.org/10.1007/978-1-4614-1472-8_3

Shen, Y., Shepherd, C., Ahmed, C.M., Shen, S., Yu, S.: Integrating deep spiking q-network into hypergame-theoretic deceptive defense for mitigating malware propagation in edge intelligence-enabled iot systems. IEEE Transactions on Services Computing **18**(3), 1487–1499 (2025) https://doi.org/10.1109/TSC.2025.3562355

Thielscher, M.: Gdl-iii: A description language for epistemic general game playing. In: IJCAI, pp. 1276–1282 (2017). https://doi.org/10.24963/ijcai.2017/177

Tweedale, J., Ichalkaranje, N., Sioutis, C., Jarvis, B., Consoli, A., Phillips-Wren, G.:

Innovations in multi-agent systems. Journal of Network and Computer Applications **30**(3), 1089–1115 (2007) https://doi.org/10.1016/j.jnca.2006.04.005

Trencsenyi, V., Mensfelt, A., Stathis, K.: Approximating Human Strategic Reasoning with LLM-Enhanced Recursive Reasoners Leveraging Multi-agent Hypergames (2025). https://arxiv.org/abs/2502.07443

Trencsenyi, V., Mensfelt, A., Stathis, K.: The Influence of Human-inspired Agentic Sophistication in LLM-driven Strategic Reasoners (2025). https://arxiv.org/abs/2505.09396

Thomas, C.K., Saad, W.: Hypergame Theory for Decentralized Resource Allocation in Multi-user Semantic Communications (2024). https://arxiv.org/abs/2409.17985

Tang, Y., Sun, J., Wang, H., Deng, J., Tong, L., Xu, W.: A method of network attack-defense game and collaborative defense decision-making based on hierarchical multi-agent reinforcement learning. Computers & Security **142**, 103871 (2024) https://doi.org/10.1016/j.cose.2024.103871

Thukkaraju, A.R., Yoon, H.J., Matsumoto, S., Ferrari, J.F., Lee, D., Ahn, M.K., Costa, P., Cho, J.-H.: Interdependent mission impact assessment of an iot system with hypergame-theoretic attack-defense behavior modeling. In: 2023 31st International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), pp. 1–8 (2023). https://doi.org/10.1109/MASCOTS59514.2023.10387597

Udupa, S., Hemida, A., Kamhoua, C.A., Fu, J.: Reactive Synthesis of Sensor Revealing Strategies in Hypergames on Graphs (2024). https://arxiv.org/abs/2412.01975

Vane, R.R.: Hypergame theory for dtgt agents. In: American Association for Artificial Intelligence (2000)

Vane, R.: Advances in hypergame theory. In: Workshop on Game Theoretic and Decision Theoretic Agents—Conference on Autonomous Agents and Multi-Agent Systems (2006)

Vane, R.R., Lehner, P.E.: Using hypergames to select plans in competitive environments. PhD thesis, George Mason University, USA (2000). https://doi.org/10.5555/932798

Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition). Princeton University Press, Princeton, New Jersey (1944)

Wan, Z., Cho, J.-H., Zhu, M., Anwar, A.H., Kamhoua, C., Singh, M.P.: Foureye: Defensive Deception based on Hypergame Theory Against Advanced Persistent Threats (2021). https://arxiv.org/abs/2101.02863

Wan, Z., Cho, J.-H., Zhu, M., Anwar, A.H., Kamhoua, C.A., Singh, M.P.: Resisting multiple advanced persistent threats via hypergame-theoretic defensive deception. IEEE Transactions on Network and Service Management **20**(3), 3816–3830 (2023) https://doi.org/10.1109/TNSM.2023.3240366

Wang, M., Hipel, K.W., Fraser, N.M.: Modeling misperceptions in games. Behavioral Science **33**(3), 207–223 (1988) https://doi.org/10.1002/bs.3830330305

Xi, B., Kamhoua, C.A.: 3. A Hypergame-Based Defense Strategy Toward Cyber Deception in Internet of Battlefield Things (IoBT), pp. 59–77. John Wiley & Sons, Ltd, USA (2020). https://doi.org/10.1002/9781119593386.ch3

Yuan, M., Cheng, Z., Ma, T.: Equilibrium analysis of distributed aggregative game with misinformation. Kybernetika **60**(6), 754–778 (2024) https://doi.org/10.14736/kyb-2024-6-0754

Yoon, H.J., Thukkaraju, A.R., Matsumoto, S., Ferrari, J.F., Lee, D., Ahn, M.K., Costa, P., Cho, J.-H.: Subjective-bayesian-network-based interdependent mission impact assessment with game-theoretic attack–defense interactions. IEEE Internet of Things Journal **12**(6), 6548–6562 (2025) https://doi.org/10.1109/JIOT.2024.3492710

Zhu, M., Anwar, A.H., Wan, Z., Cho, J.-H., Kamhoua, C.A., Singh, M.P.: A survey of defensive deception: Approaches using game theory and machine learning. IEEE Communications Surveys & Tutorials **23**(4), 2460–2493 (2021) https://doi.org/10.1109/COMST.2021.3102874